Check for updates

Emerging materials intelligence ecosystems propelled by machine learning

Rohit Batra[1], *Le Song*[2] *and Rampi Ramprasad*[3]✉

Abstract | The age of cognitive computing and artificial intelligence (AI) is just dawning. Inspired by its successes and promises, several AI ecosystems are blossoming, many of them within the domain of materials science and engineering. These materials intelligence ecosystems are being shaped by several independent developments. Machine learning (ML) algorithms and extant materials data are utilized to create surrogate models of materials properties and performance predictions. Materials data repositories, which fuel such surrogate model development, are mushrooming. Automated data and knowledge capture from the literature (to populate data repositories) using natural language processing approaches is being explored. The design of materials that meet target property requirements and of synthesis steps to create target materials appear to be within reach, either by closed-loop active-learning strategies or by inverting the prediction pipeline using advanced generative algorithms. AI and ML concepts are also transforming the computational and physical laboratory infrastructural landscapes used to create materials data in the first place. Surrogate models that can outstrip physics-based simulations (on which they are trained) by several orders of magnitude in speed while preserving accuracy are being actively developed. Automation, autonomy and guided high-throughput techniques are imparting enormous efficiencies and eliminating redundancies in materials synthesis and characterization. The integration of the various parts of the burgeoning ML landscape may lead to materials-savvy digital assistants and to a human–machine partnership that could enable dramatic efficiencies, accelerated discoveries and increased productivity. Here, we review these emergent materials intelligence ecosystems and discuss the imminent challenges and opportunities.

Using available data or observations to formulate decisions, conclusions and even theories is not new. In fact, this paradigm has been in existence during the entire course of human (or animate) history. A growing child, when continuously exposed to a variety of inputs and incentives, progressively learns to form strategies and hypotheses to deal with the world. Nature sets the asymptotic limit of the level of intelligence a species may be able to achieve, and nurture provides the exposure and avenues to achieve the asymptotic limit.

Reproducing experience-based learning and decision-making in machines, thus, producing 'artificial intelligence' (AI), has led to a lot of excitement (and expectations) in recent decades. Two aspects make the contemporary situation truly unique though. First, the size of available data, at least in some specific contexts, is enormous, of the same magnitude or much larger than that encountered by natural cognitive systems. This situation has spawned new ways of representing, processing and learning from data. Second, in contexts where data are already immense, there is also a constant tsunami of new data. Learning models — or 'machine learning' (ML) models — thus, have the opportunity to continuously learn from the flux of incoming data, and, more importantly, demand new data in regions of sparse knowledge. The machine, much like a growing child, can progressively improve in intelligence or capability in a self-directed or autonomous manner. AI, at least in specific data-rich contexts, is very much a reality. Classic examples have emerged in diverse domains, such as in e-commerce, computer games, autonomous driving and, also, unfortunately, in human behaviour (for instance, voter) manipulation.

The above developments are rapidly beginning to impact science and engineering, both in terms of added value and of expectations of what might be achieved. Although much of materials science is not (yet) in a data-rich situation, AI tools and their far-reaching

[1] *Center for Nanoscale Materials, Argonne National Laboratory, Lemont, IL, USA.*

[2] *Computational Science & Engineering, Georgia Institute of Technology, Atlanta, GA, USA.*

[3] *School of Materials Science & Engineering, Georgia Institute of Technology, Atlanta, GA, USA.*

✉*e-mail: rampi.ramprasad@mse.gatech.edu*

https://doi.org/10.1038/s41578-020-00255-y

potential are shaping a veritable materials intelligence ecosystem. Thanks to several independent developments, it is becoming a mainstream belief that materials data, if handled and harnessed appropriately, may be used to accelerate materials development and discovery, at a pace and scale that has never been seen before.

By definition, data-driven efforts start with data. In an attempt to respond to this need, materials repositories targeting several classes of materials have burgeoned. These repositories are populated with available experimental or computational data, and progressively augmented with data that are published, shared or created in a targeted manner (for example, via high-throughput experiments or computations). An area of active inquiry and growth is the development of sustainable and scalable protocols to create diverse and comprehensive data collections in a systematic and organized manner, for example, using active learning to drive experiments or computations, or natural language processing (NLP) to continuously capture data from the scientific literature and patents.

The next important component of a data-driven effort is the representation of the data themselves, so that they are converted to a machine-readable form. Materials data repositories provide datasets that are typically an enumeration of a variety of materials that fall within a well-defined chemical subclass (including some details on the manner in which the experiments — physical or computational — were performed) and relevant measured (or computed) properties or synthesis steps adopted. Converting such data to a machine-readable form involves numerically representing the materials and other relevant details. This step, referred to as 'fingerprinting', intrinsically depends on the context or application, and leads to a spectrum of numbers, or descriptors, that capture key attributes of the material or process, which may be either 'handcrafted' or automatically generated.

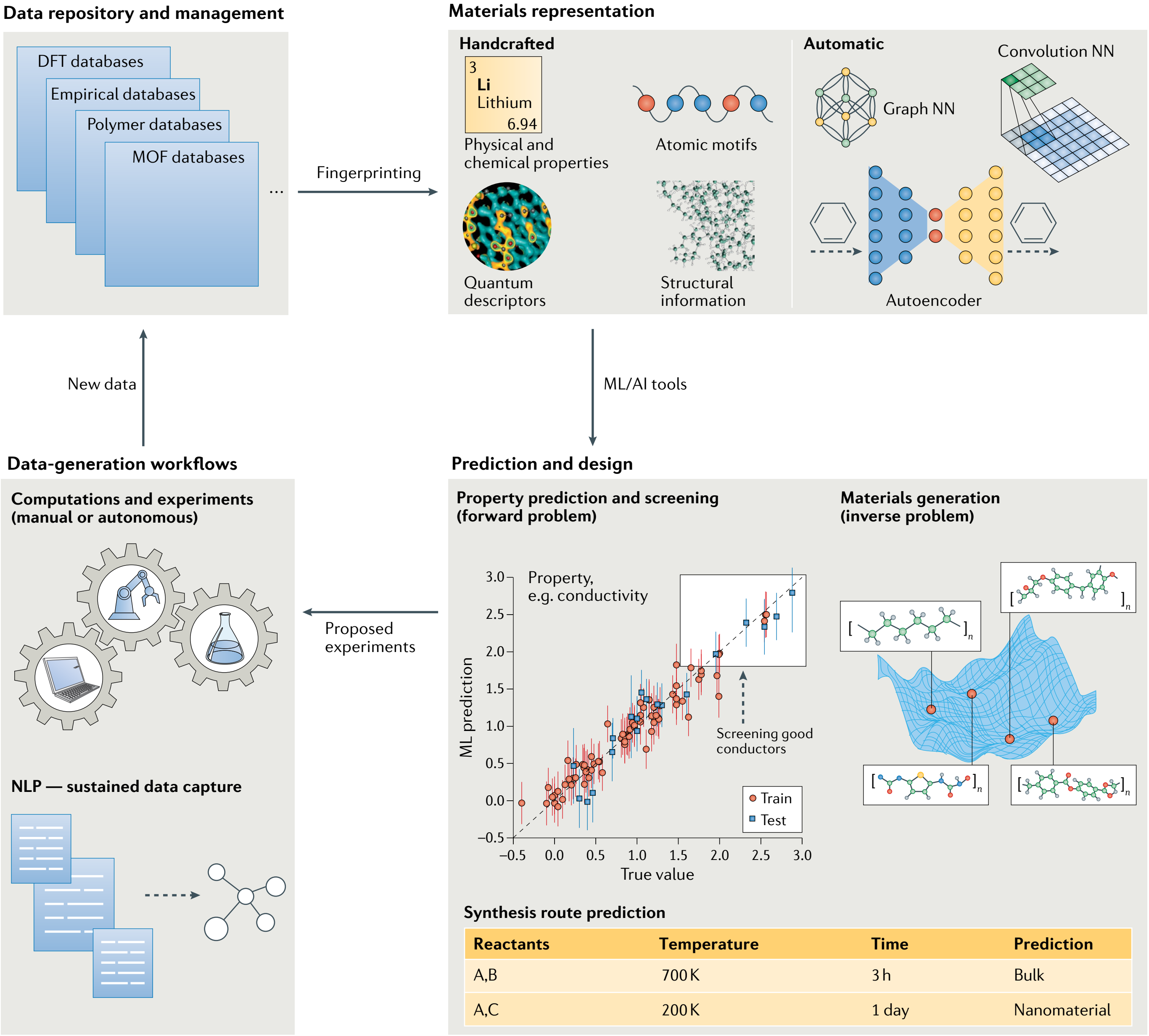
With a machine-readable representation at hand, and, of course, a large and diverse enough dataset, progress can be made to detect patterns in the data (via unsupervised learning) or to make predictions of properties or synthesis recipes for new materials (via supervised learning). Unsupervised learning is a problem that involves just the machine-readable materials representations and looks for similarities and differences between the various cases. It learns patterns in the data, such as clusters and extremum data points. Supervised learning, by contrast, involves a training process that establishes a mapping between the representations and the properties or synthesis outcomes, leading to surrogate models of predictors. The last several years have seen enormous progress on these fronts for a variety of materials classes and phenomena. A dizzying spectrum of old and new algorithms have been utilized, ranging from linear regression and nonlinear methods (kernel-based or Gaussian-process-based) to decision trees and neural networks (NNs; shallow, deep or convolutional)[1]. More recently, emphasis has shifted to strategies for solving the 'inverse problem', that is, the enumeration of materials or process/synthesis designs that are expected to meet a property or performance target with high probability. Inspired by capabilities developed by the image and video generation community, generative models — along with traditional approaches, such as active-learning and evolutionary algorithms — are making inroads into materials discovery. These models enable the search of superior materials in a proxy latent space, allowing the use of powerful optimization approaches for materials design. Screened candidate materials are then evaluated through experiments and computations, and become part of the ever-growing data repositories.

The computational and physical laboratory infrastructures are also transforming, owing to the integration of AI tools[2]. On the computational side, AI agents that learn the input–output behaviour of simulation software can be several orders of magnitude faster than traditional approaches[3,4]. Likewise, materials synthesis and characterization facilities are beginning to see dramatic improvements in efficiencies due to the integration of ML capabilities (which provide autonomy) and the incorporation of robotic control (which imparts automation)[5]. Although these AI agents need to be further nurtured, the natural evolution of such human–machine partnerships may lead to materials-savvy digital assistants that will continuously and autonomously learn aspects of materials science and engineering.

Here, we review some of the mature and emergent key components of the materials intelligence ecosystem (FIG. 1). We pay special attention to the protocols for data acquisition and management, context-dependent representation of data, transformation of data to surrogate predictive models and knowledge, and emerging strategies for solving inverse problems[6] that can autonomously drive a materials laboratory. Throughout, we provide examples of materials innovations that have resulted from the infusion of AI ideas within materials science and engineering, and highlight the imminent challenges and opportunities.

Data generation, acquisition and management

Powered by the Materials Genome Initiative, the general data-management policies enforced by funding agencies and the recent awareness within the materials community of the positive impact of data sharing and dissemination, several efforts to build materials databases have blossomed. It is fair to say that a vast majority of materials databases have grown organically with proactive discussions on management standards, policies and on the associated challenges. Consequently, FAIR (findable, accessible, interoperable and reusable) data principles that provide guidelines for scientific data management and stewardship have been put forth[7]. A myriad of databases, both empirical and computational, spanning a large variety of materials properties — including structures, formation energetics, thermodynamic phase diagrams, electrical and mechanical properties — across different material classes — metals, ceramics, alloys, glasses, 2D materials and nanocomposites — have become available. A few notable examples are presented in TABLE 1. Importantly, several of these datasets are coupled with data visualization or search tools, or are accessible through an application programming interface (API), thereby, allowing easy and quick access, and supporting the acceleration of materials discovery. The success and



Reactants	Temperature	Time	Prediction
A,B	700 K	3 h	Bulk
A,C	200 K	1 day	Nanomaterial

Fig. 1 | **Materials intelligence ecosystems.** Materials intelligence ecosystems consist of four interconnected key components. First, online repositories provide user-friendly access to a range of materials data (including properties, structure, composition and microstructural images). Second, manually constructed or machine-based representation methods transform the available materials data into a numerical format, making it amenable to artificial intelligence (AI) and machine learning (ML) tools. Third, surrogate or generative models use the processed materials data to screen or generate new candidates with desired properties. The proposed candidates can either be new materials with superior properties, or potential synthesis routes (reactants, precursors) or processing conditions (annealing, calcination temperature) to obtain target materials. Lastly, the identified 'best' candidate is validated through experimental or computational tests, with the outcomes appended to the existing repositories. More recently, tools based on natural language processing (NLP) have been employed to directly mine the literature corpus, producing accessible and user-friendly materials data. DFT, density functional theory; MOF, metal–organic framework; NN, neural network. The image for the convolution NN is courtesy of V. Dumoulin and F. Visin. The image for property prediction and screening is reprinted with permission from REF.[62], Elsevier. The graph for the graph neural network is reprinted with permission from REF.[107], APS.

potential impact of such repositories can be enhanced if materials data infrastructure challenges, such as the scarcity of open or common data standards and formats, the lack of e-collaborative platforms and tools, and the insufficient reward of data contributors, are addressed[8].

To make data generation (and augmentation of existing repositories) a sustainable, painless and efficient enterprise, steps are being taken, as we discuss below.

Design of experiments. We start by summarizing a traditional approach: choosing a small subset of cases to perform experiments on from a potentially large candidate set, with the constraint that the former is representative of the latter. Experiments (empirical or computational) in materials science usually involve exploring or optimizing a large number of parameters; for example, organic photovoltaic materials require

Table 1 | **Notable materials repositories**

Name	Material types	Source	No. of entries	Access
NIST ICSD[231]	Inorganic	Empirical	210,000	License
Pauling File[232]	Inorganic	Empirical	156,274	Open
PoLyInfo[233]	Polymers	Empirical	334,738	Open
Cambridge Structural Database[234]	Organic, MOFs	Empirical	>1 million	Open/license
MatWeb[235]	Inorganic, organic	Empirical	135,000	License
Total Metals[236]	Metals	Empirical	350,000	License
INTERGLAD[237]	Glasses	Empirical	350,000	License
Mindat[238]	Minerals	Empirical	5,500	Open
ASM Databases & Handbooks[239]	Alloys	Empirical	–	License
American Mineralogist Crystal Structure Database[240]	Minerals	Empirical	–	Open
Citrination[241]	General materials	Empirical, computational	350,000	Open
FIZ Karlsruhe ICSD[242]	Inorganic	Empirical, computational	>210,000	License
ChemSpider[243]	Organic	Empirical, computational	81 million	Open
MatNavi NIMS Databases[244]	General materials	Empirical, computational	–	Open
NIST Materials Data Repository[245]	General materials	Empirical, computational	–	Open
NanoMine[246]	Polymer nanocomposites	Empirical, computational	–	Open
SpringerMaterials Databases[247]	General materials	Empirical, computational	–	License
Crystallography Open Database[248]	General materials	Computational	451,943	Open
Materials Project[249]	Inorganic	Computational (DFT)	>120,000	Open
OQMD[250]	Inorganic	Computational (DFT)	637,644	Open
AFLOW[251]	Inorganic	Computational	3,225,000	Open
Jarvis[252]	Inorganic	Computational (DFT)	>30,000	Open
f-Electron Structure Database[253]	Inorganic	Computational	28,000	Open
CatApp[254]	Catalysis	Computational (DFT)	1,054	Open
NOMAD[255]	General materials	Computational (quantum)	–	Open
Novamag[256]	Rare-earth magnets	Computational	–	Open
CALPHAD Databases[257]	Inorganic	Computational	–	License
Computational Materials Repository[258]	Inorganic	Computational	–	Open
MaterialsWeb[259]	2D/3D inorganic	Computational	–	Open

For some databases, the information on the number of entries is not available and, for others, it changes too fast to provide an accurate value. DFT, density functional theory; MOFs, metal–organic frameworks.

optimizing the donor-to-acceptor ratio, the thickness of the heterojunction layer, the processing additives and the spin-casting speed. In most scenarios, the experimental budget is too limited to ensure adequate sampling of the entire parameter space. The design of experiments (DOE) approach is utilized to sample a large, multidimensional parameter space in a rational manner with minimal budget[9,10]. Naturally, the objective is to adopt a space-filling sampling approach that provides information over the entire parameter space. But, as illustrated in FIG. 2a, a grid-based uniform sampling approach may lead to selecting the same parameter value multiple times, which can be avoided using an approach called the Latin hypercube design (LHD)[11]. In this approach, the design space is split into grid-based hypercubes (bins), and the points are sampled such that no two points have the same value for any of the design parameters. An exemplary LHD schema is illustrated in FIG. 2a, with the shaded regions highlighting the absence of any duplicate selection. Further, to distribute the points across the design space, a 'maximin' criterion is imposed that maximizes the distance between two sampled points, while simultaneously minimizing the number of points having similar distance values. More advanced versions of LHD include sliced-LHD[12], MaxPro-LHD[13] and Pareto optimal LHD[14], among others[15], which may be used to create batches of physical or computational experiments that are simultaneously sparse and diverse.

Despite its usefulness, the use of LHD in materials science has been limited. Notable examples include optimizing synthesis parameters for the design of efficient organic photovoltaics[10] and determining simulation parameters to generate a mechanical-property database using finite-element simulations[16,17]. Nonetheless, the use of LHD in materials science is expected to rise, especially for the initial generation of diverse databases to be later expanded and exploited with other ML methods.

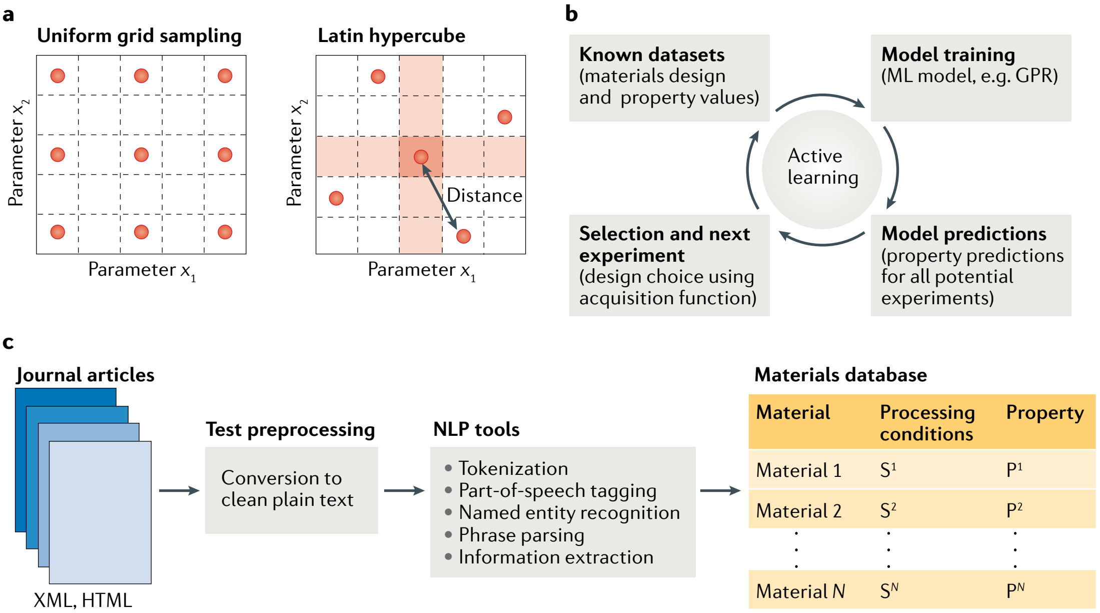
Active learning or Bayesian optimization. The DOE approach may lead to an economical but diverse selection of cases to perform experiments on, but decisions must be done ahead of time, and cannot be changed during the course of the experiments. An alternative approach is based on Bayesian optimization or active learning[18,19] (FIG. 2b), in which successive experiments are decided based on the outcomes of past experiments. An initial set of experiments is performed, a model (for example, one based on Gaussian processes) is fit to the available data, predictions are made on a large set of potential experiments and the next experiment is selected based on some criteria related to the predictions. In other words, decisions on what to do next are made not just by considering the possible experimental design space (as in the DOE approaches) but also based on the available experimental outputs, that is, the materials property space. Subsequently, the candidate for which the experiment is performed is added to the training set, and the procedure is repeated iteratively to build the dataset in a targeted manner.

The criterion used for selecting the next experiment — commonly referred to as the acquisition function — is critical. Methods such as Gaussian process regression (GPR)[18], which can fit a model to the data from experiments that have already been performed, can provide both a prediction and an uncertainty on the prediction for new cases. One may select the next experiment based purely on the uncertainty values. Large uncertainties indicate a region in the feature space with poor knowledge. Acquisition functions defined in this manner will lead to exploration and visits to regions of sparse knowledge or data.

Alternatively, the objective may be to find the best material candidate with the desired property values. In such cases, the acquisition function is defined based on the model predictions themselves, that is, the next experiment is selected based on how close its predicted property value is to the desired value. This strategy is called exploitation.

Bayesian optimization combines the predicted value and its uncertainty in a single acquisition function to select the next location for measurement, hence, balancing exploration and exploitation[20]. The acquisition function can be designed in many different ways, including maximum probability of improvement, maximum expected improvement[21] and Thompson sampling[22]. These methods are collectively classified as Bayesian optimization or active learning, and are increasingly becoming a vital part of the materials science ML portfolio to plan experiments[23,24].

As an example, Bayesian optimization was used to guide the discovery of piezoelectrics with the ABO_3 perovskite structure with large electrostrains through suitable quantities of dopant substitution at the A-site and the B-site[25]. Starting with experimental measurements



Material	Processing conditions	Property
Material 1	S^1	P^1
Material 2	S^2	P^2
⋮	⋮	⋮
Material *N*	S^N	P^N

Fig. 2 | **Strategies for materials data generation and acquisition. a** | Schematic illustration of the design of experiments approach to select the set of experiments to perform. Whereas traditional grid-based selection leads to redundant experiments performed at the same parameter values, design of experiments strategies (such as the Latin hypercube) cover the parameter space with minimal redundancy and minimal experimental budget. **b** | Active learning and Bayesian optimization scheme for iterative data generation and model improvement. **c** | Steps involved in the extraction of materials data from the scientific literature using natural language processing (NLP). GPR, Gaussian process regression; ML, machine learning.

of only 61 cases from a total of 605,000 possible compositions, a bootstrapping method was used to generate statistically equivalent training datasets of 1,000 samples. An ensemble of 1,000 ML models consisting of polynomial fits, support vector machines with a linear and radial-based kernel functions, and gradient tree boosting was then used to obtain predictions (with means and variances) for all candidate compositions. These predictions were plugged into various acquisition functions (exploration, exploitation, trade-off) to identify candidates for the next set of measurements. This active-learning scheme was iterated for five loops and resulted in the identification of $(Ba_{0.84}Ca_{0.16})(Ti_{0.90}Zr_{0.07}Sn_{0.03})O_3$, which has an electrostrain of 0.23%, compared with the highest value of 0.15% in the training data. Similar works using Bayesian optimization for accelerating materials discovery include the design of high-strength, high-entropy alloys[26], shape-memory alloys with low thermal hysteresis ($Ti_{50.0}Ni_{46.7}Cu_{0.8}Fe_{2.3}Pd_{0.2}$)[23] and piezoelectrics with vertical morphotropic phase boundary ($(Ba_{0.5}Ca_{0.5})TiO_3$-$Ba(Ti_{0.7}Zr_{0.3})O_3$)[27]; review articles on this topic are also available[28,29].

An important by-product of Bayesian optimization is the materials data generated as part of the new measurements or computations. Thus, besides providing an efficient strategy to find desirable materials, it also provides an opportunity to generate new materials data that can be added to existing materials repositories. However, an important caveat to bear in mind is that the benefits of Bayesian optimization are strongly influenced by the model parameters and the definition of the acquisition function, as underscored by the benchmark study in REF.[30]. Furthermore, active-learning strategies can only be used when new experiments for the identified candidates can be conducted in a reasonable time. Therefore, Bayesian optimization may be inefficient or expensive when the parameter search space is enormous (for example, in the game of Go or in the search for the lowest-energy configuration in a high-entropy alloy). This is due to the computational complexity of optimizing in a high-dimensional space where the optimization objective is non-convex. Nonetheless, Bayesian optimization is one of the most successful techniques in the materials intelligence ecosystem, especially for problems involving costly experiments or relatively small datasets ($\lesssim$200 points).

NLP. Computational and physical experiments can be performed in an organized manner to systematically create materials data, but an immense amount of untapped data are already in the publication and patent literature, and this content is increasing exponentially. A great opportunity exists to extract and capture structured data from such corpora, in a sustainable and automated manner, for regular ingestion by data repositories. Software pipelines are being developed to mine natural-language texts, scientific figures and tables to find interesting materials trends, suggest future materials and extract materials property and synthesis information from millions of documents.

A collection of NLP tools, parsing algorithms and journal APIs are necessary to mine relevant materials data in an autonomous and machine-processable format (FIG. 2c). The first step is the retrieval of relevant journal articles in HTML, XML or PDF format through the use of DOIs, journal APIs and Crossrefs. Next, some basic text processing is performed to clean and convert the document in raw plain text. Following this, a series of NLP operations[31] are conducted: tokenization, the process of converting plain text into contextual tokens, which broadly correspond to individual words and punctuation that make a meaningful building block (such as ‘thermoelectric’, ‘$BaTiO_3$’, ‘$Cu_{0.3}Al_{0.7}$’ or ‘sintering’); part-of-speech (POS) tagging, which involves assigning a tag to each token to describe its syntactic function (such as noun, verb or adjective); named entity recognition (NER), which identifies the key entity (whether it is a material, property or process) for which description is available in the text; phrase parsing, which transforms a sequence of tagged tokens into a tree structure that represents the syntactic structure of each sentence (using predefined syntactic rules); and, finally, information extraction, which involves post-processing to resolve data interdependencies throughout the different sections of a document (abstract, methods, results) and combining them into a single structured record for each unique chemical entity identified within the document. In the end, the extracted records are compiled into a monolithic database, which can be programmatically queried to reveal hidden trends across huge materials domains using data analytics and visualization techniques.

The POS tagging and NER are both crucial components of the NLP pipeline, and are accomplished using both unsupervised and supervised ML models. The former usually entails numerical or vectorial representation of words (or tokens), referred to as ‘word embeddings’, using unsupervised algorithms, such as GloVe[32] or Word2vec[33]. The underlying principle is to assign high-dimensional vectors (or embeddings) to all words in a text corpus in such a manner that preserves their syntactic and semantic relationships. This is achieved using information about the co-occurrences of the words in a text corpus such that words used within the same context are mapped closer (cosine distance) to each other within the numerical latent space. For example, the embedding of ‘aluminum’ would be closer to ‘metal’ than to ‘non-metal’. Moreover, it was demonstrated[34] that a wealth of materials science knowledge is also captured by the latent space of word embeddings, including the underlying structure of the periodic table and different structure–property relationships (such as melting temperatures, electronegativities and formation energies) in materials. For instance, the reduced 2D representation of the word embeddings of elements correctly segregated alkali metals, alkaline earth metals, transition metals, actinides, halogens and noble gases in different clusters, similar to the periodic table. The information contained within word embeddings is also evident from the capability of the ML model to answer analogies: ‘NiFe’ is to ‘ferromagnetic’ as ‘IrMn’ is to ‘?’, the model correctly answers ‘antiferromagnetic’. More impressively, these word embeddings have been shown to be capable of recommending materials for future discovery by exploiting the complex relationships learned from the massive body of scientific literature. Taking the example of thermoelectrics, it was shown that, by simply ranking the materials based on their cosine similarity with the word ‘thermoelectric’, it is possible to not only recover known thermoelectrics but also recommend new materials systems (such as Li_2CuSb and $CuBiS_2$) for future studies. Interestingly, this study demonstrated that the latent space distances learned by the unsupervised model can also be used to make (qualitative) property predictions, which is usually formulated as a supervised learning problem.

As part of the NLP pipeline, the learned word embeddings are also utilized to train supervised POS tagging or NER models based on manually constructed training sets of annotated tags for words appearing in a small subset of journal articles. Once trained, the POS tagging or NER models can be applied to countless journal articles in an automated manner and extract key materials information. In fact, NLP has been successfully used to extract materials property and synthesis information from different materials domains, including Curie and Néel temperatures for magnetic materials[35], synthesis conditions for metal oxides and zeolites[36], and to gain insights for inorganic materials synthesis planning[37,38]. In the last example, the researchers scanned through over 12,000 synthesis articles from a pool of half a million journal articles to compile synthesis outcomes of calcination, sol–gel, hydrothermal or solid-state routes for inorganic materials. Once compiled, the dataset was analysed to find trends. For instance, increasing elemental complexity from binaries to ternaries to pentanaries required higher calcination temperatures, exceeding 400 °C, for the synthesis of bulk and nanostructured materials; this is expected, because multicomponent systems require interdiffusion of multiple species, whereas binaries can be formed from the oxidation of carbonate or hydroxyl groups. By contrast, most hydrothermal reactions were found to be conducted at much lower temperatures of 150–200 °C for 12 or 24 h, irrespective of the number of elements involved. The compiled synthesis database could also be trained in a supervised learning fashion to predict synthesis outcomes. For example, given input synthesis conditions, the classification models predicted the formation of tetragonal (ferroelectric) $BiFeO_3$, 2D CdS and nanotube titania[37]. For the last problem, 27 synthesis variables (such as annealing temperature and drying time) obtained from 22,065 journal articles were used to develop a decision tree model, which revealed that NaOH concentration and hydrothermal temperature are the two key parameters driving the formation of nanotube titania. These ideas have been further extended to predict the synthesis conditions themselves, that is, precursors, annealing temperatures and time, necessary to produce the target materials using generative models[38]. Strategies to overcome the challenges encountered in clearly distinguishing the roles of different chemical entities as, for example, reagents, targets or media, have been proposed[39]. Another interesting work based on NLP is a machine-generated review on Li-ion batteries that was extracted from over 150 journal articles[40].

Although the field of NLP for materials data retrieval is relatively new, its popularity is expected to grow exponentially, owing to the availability of powerful open-source codes capable of retrieving information of different types (properties, processing, synthesis) and from different formats (text, tables, figures). However, a few challenges remain. The POS tagging relies on the labels provided by domain experts, but, depending on the material class, the labelling process can be tedious, time-consuming and incomplete, resulting in poor NLP performance; this is the case for polymers, for example. This problem becomes even more severe when relatively complex information needs to be extracted, such as a sequence of multiple reactions at different time intervals. Furthermore, the NLP models can be improved only if they can interact with domain experts in an active manner, continuously requesting new labels or feedback. Lastly, an important source of materials data is images and tables, for which the performance of NLP models can be improved significantly. Given the challenges in figure parsing, currently, most works have either resorted to manual methods or restricted their focus to limited domains with strong assumptions[41–43]. However, some progress has been made in extracting the plotted data from a figure and correctly associating it with the legend entries using a novel graph-based reasoning approach coupled with a deep-learning-based similarity metric[44]. NLP and computer-vision techniques have also been used to understand diagrams and text to solve Boolean-satisfiability-style geometry problems[45]. Other works[46,47] have extended this approach to obtain axiomatic knowledge of geometry and to solve geometry problems.

Basic ML algorithms for materials scientists

In this section, we frame the problem of learning from data and elaborate on several algorithms that gained prominence within the materials science community in the last decade or so. We assume that a well-curated dataset is already available. An example dataset may be an enumeration of a variety of materials that fall within a well-defined chemical class of interest (the input) and a relevant measured or computed property (the output). The (supervised) learning problem is then defined as simply establishing a generalizable mapping between material and property (or input and output).

All data-driven strategies that attempt to address this problem are composed of two distinct steps. The first step is to represent numerically the various input cases (or materials) in the dataset. At the end of this step, each input has been reduced to a string of numbers (or ‘fingerprints’, FIGS 3, 4a). The second step establishes a mapping between the fingerprinted input and the target property, and is entirely numerical. Below, we review both the fingerprinting and learning aspects. We note in passing that fingerprinting requires domain knowledge and creativity, with no single well-defined path, whereas for several ML algorithms, well-organized open-source libraries, such as sklearn[48] and TensorFlow[49], are available for the community to exploit.

Materials representation. The choice of the numerical representation can be effectively accomplished only with adequate knowledge of the problem and goals (that is, domain expertise), and typically proceeds in an iterative manner by duly considering aspects of the material that

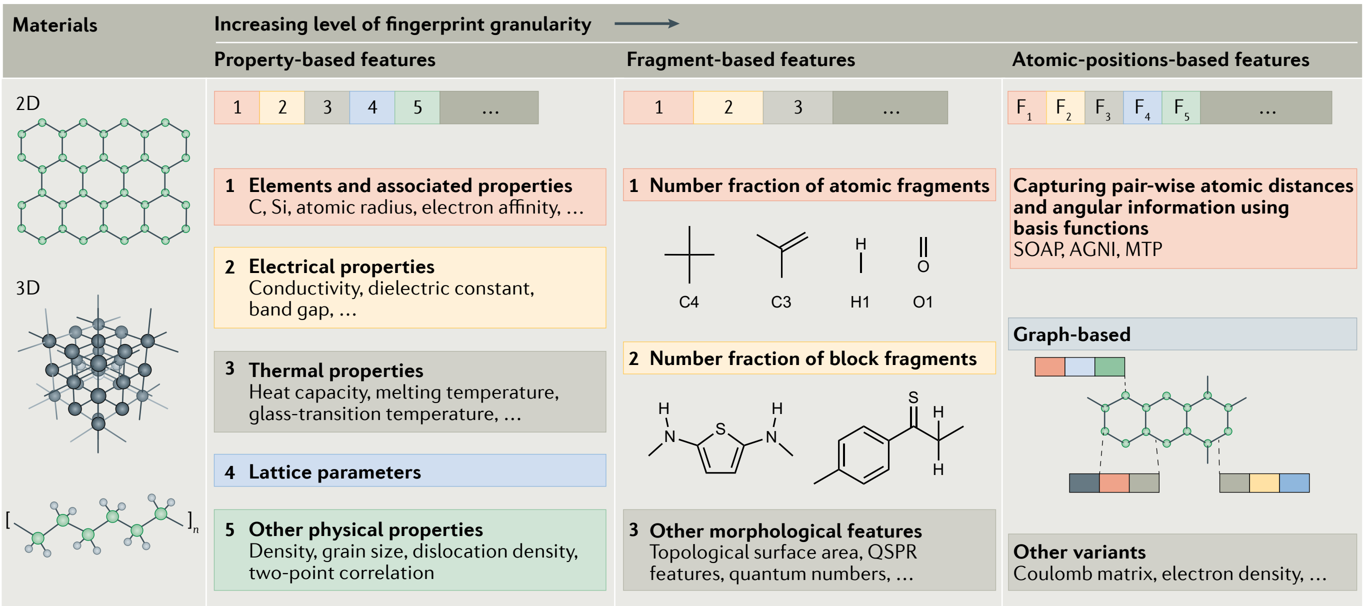

Materials
Increasing level of fingerprint granularity
Property-based features
Fragment-based features
Atomic-positions-based features
2D
3D
1 2 3 4 5 …
1 Elements and associated properties
C, Si, atomic radius, electron affinity, …
2 Electrical properties
Conductivity, dielectric constant, band gap, …
3 Thermal properties
Heat capacity, melting temperature, glass-transition temperature, …
4 Lattice parameters
5 Other physical properties
Density, grain size, dislocation density, two-point correlation
1 2 3 …
1 Number fraction of atomic fragments
C4
C3
H1
O1
2 Number fraction of block fragments
3 Other morphological features
Topological surface area, QSPR features, quantum numbers, …
F1 F2 F3 F4 F5 …
Capturing pair-wise atomic distances and angular information using basis functions
SOAP, AGNI, MTP
Graph-based
Other variants
Coulomb matrix, electron density, …

Fig. 3 | **Materials fingerprinting.** The schematic shows a hierarchy of materials fingerprinting methods with increasing levels of granularity. At the highest level are the features based on general materials properties, such as density, grain size or attributes of the constituting elemental species (such as valency or atomic radius). Features based on the presence or absence of certain atomic fragments or other local morphological descriptors form the intermediate level. At the finest level, the fingerprint numerically represents information about the atomic positions, associated bond types or even electronic-charge-density distribution. Generally speaking, the finer the fingerprint, the higher the expected model accuracy and the larger the dataset required to train the machine learning model. AGNI, adaptive generalizable neighbourhood informed; MTP, moment tensor potentials; QSPR, quantitative structure–property relationships; SOAP, smooth overlap of atomic potentials. The image of the 3D structure is adapted from REF.[3], CC BY 4.0.

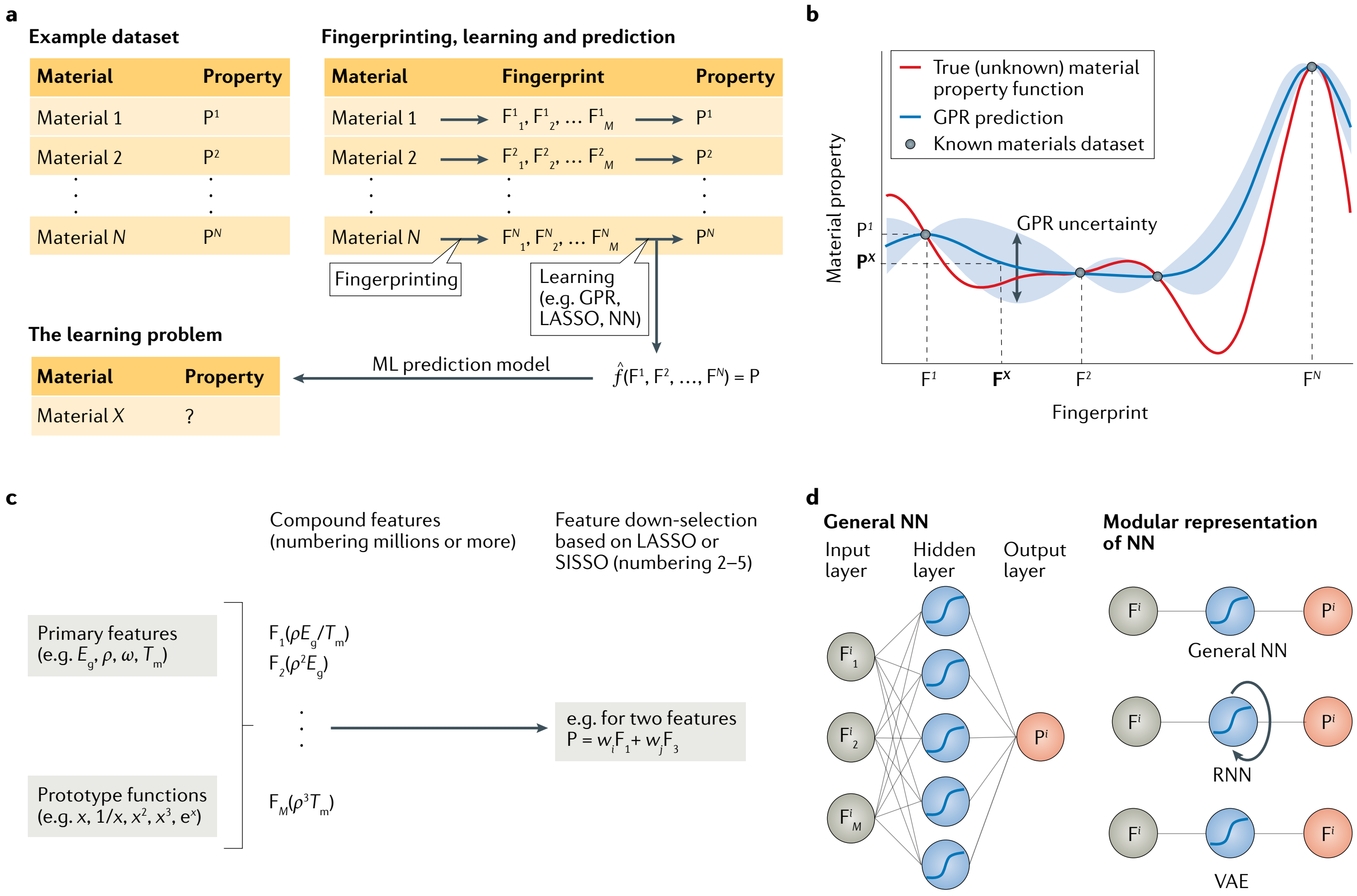
the target property may be correlated with. Given that the numerical representation serves as the proxy for the real material, it is also referred to as the fingerprint of the material or its descriptor (in ML parlance, it is also referred to as the feature vector).

Depending on the problem and the accuracy requirements of the predictions, the fingerprint can be defined at varying levels of granularity, as shown in FIG. 3. For instance, if the goal is to obtain a high-level understanding of the factors underlying a complex phenomenon — such as the mechanical or electrical strength of a material or its catalytic activity — and prediction accuracy is not critical, then the fingerprint may be defined at a gross level, in terms of the general attributes of the atoms the material is made up of, other potentially relevant properties (such as the band gap) or higher-level structural features (such as the typical grain size). By contrast, if the goal is to predict specific properties at a reasonable level of accuracy across a wide chemical space — such as the dielectric constant of an insulator or the glass-transition temperature of a polymer — the fingerprint may have to include information pertaining to key atomic-level structural fragments that may control properties. If extreme (chemical) accuracy in predictions is demanded — such as electronic charge density, total energies, atomic forces, precise identification of structural features, space groups or phases — the fingerprint has to be fine enough to encode details of atomic-level structural information with sub-angstrom resolution. Several examples of learning based on this hierarchy of fingerprints or descriptors have been discussed in the past[3,50,51].

The general rule of thumb is that the finer the fingerprint, the greater the expected accuracy, and the more laborious, data-intensive and less conceptual the learning framework. As we discuss below, this opens up the possibility of automatically learning and determining the fingerprints in such problems. Finally, regardless of the application, the fingerprints should be invariant (or covariant in some scenarios) to certain transformations of the system, such as rotation, translation and permutation of like members of the system.

Regression and multifidelity learning. Once a fingerprinting scheme is selected and all materials in a database have been fingerprinted, the regression problem is to find a function $\hat{f}(\,.\,)$ that takes a material's fingerprint as input and returns its associated target property as output (FIG. 4a). The learned mapping $\hat{f}(\,.\,)$ is called the surrogate or the ML learning model, whereas the materials database used to fit the model $\hat{f}(\,.\,)$ is called the training data. The advantage of the learned ML model is that it can be used to make quick property predictions for new materials (denoted as X in FIG. 4a) by first fingerprinting them and then simply applying the learned ML model $\hat{f}(\,.\,)$ to obtain their property estimates. Obviously, using ML only makes sense if fingerprinting and evaluating the mapping function for the new case is significantly faster than directly making the property measurement.

Several strategies for arriving at the mapping function exploit different regression algorithms, including linear regression, kernel ridge regression (KRR), GPR, decision trees, random forest, support vector machine



Example dataset

Material	Property
Material 1	P^1
Material 2	P^2
⋮	⋮
Material *N*	P^N

Fingerprinting, learning and prediction

Material	Fingerprint	Property
Material 1 →	$F^1_1, F^1_2, \dots F^1_M$ →	P^1
Material 2 →	$F^2_1, F^2_2, \dots F^2_M$ →	P^2
⋮	⋮	⋮
Material *N* →	$F^N_1, F^N_2, \dots F^N_M$ →	P^N

The learning problem

Material	Property
Material *X*	?

a
Fingerprinting
Learning (e.g. GPR, LASSO, NN)
ML prediction model
$\hat{f}(F^1, F^2, \dots, F^N) = P$
b
True (unknown) material property function
GPR prediction
Known materials dataset
Material property
P^1
P^X
GPR uncertainty
F^1
F^X
F^2
F^N
Fingerprint
c
Compound features (numbering millions or more)
Feature down-selection based on LASSO or SISSO (numbering 2–5)
Primary features (e.g. E_g, ρ, ω, T_m)
Prototype functions (e.g. x, 1/x, x^2, x^3, e^x)
$F_1(\rho E_g/T_m)$
$F_2(\rho^2 E_g)$
$F_M(\rho^3 T_m)$
e.g. for two features $P = w_i F_1 + w_j F_3$
d
General NN
Input layer
Hidden layer
Output layer
F^i_1
F^i_2
F^i_M
P^i
Modular representation of NN
F^i
P^i
General NN
RNN
VAE

Fig. 4 | **The general learning problem in materials science and its solution using common machine learning techniques.** **a** | An example materials dataset, with the statement of the learning problem, and a schematic representation of the creation of a surrogate prediction model via the fingerprinting and learning steps. **b** | Gaussian process regression (GPR) mapping the fingerprint space to the materials property space. The GPR can provide a quick property estimate for a new material with fingerprint F^x, along with the prediction uncertainty (shaded region). **c** | Feature selection using the least absolute shrinkage and selection operator (LASSO) to reveal simple functional solutions of the learning problem. **d** | Illustration of a neural network (NN) that takes material fingerprints as input, transforms them into a more abstract representation in the hidden layer and, finally, maps them to the target output property value. Modular representations of a basic NN, recurrent NN (RNN) and of a variational autoencoder (VAE) highlighting the information flow in each case are also shown. ML, machine learning; SISSO, sure independence screening and sparsifying operator. Panel **a** is adapted from REF.[3], CC BY 4.0.

regression and NNs. These algorithms mostly vary in the type and complexity of the function space that is searched to fit the training data. Another key aspect is the uncertainty in the model prediction, as illustrated in FIG. 4b for the case of GPR. These uncertainties are often used to quantify the confidence in the model prediction, and, thus, are useful to drive the active-learning approach by evaluating the acquisition function (see the section on active learning).

The fingerprint definition and the regression algorithm together define the quality and performance that the developed ML models can achieve. If the available target property (output) varies smoothly and is not highly nonlinear in the fingerprint space, a relatively simple mapping function can achieve good performance. Thus, in many cases, rather than working in the original fingerprint space, the learning problem is posed in a modified kernel space, wherein the transformed fingerprint follows a much simpler trend with the target property, allowing the ML model to reach better accuracy. Another important aspect is the chosen complexity of the mapping function (or the regression algorithm), which dictates its ability to fit the given training data; highly nonlinear functions, owing to their flexible nature, can provide a better fit to the data. However, caution should be exercised, as they may not generalize well for new materials (test data) and could result in overfitting, which means that the learned ML model performs well only for the materials included in the training data and provides arbitrary or poor predictions for materials outside the training set. Such behaviour can be avoided using a regularization scheme during model training, a strategy almost always exploited in practice. Regularization can be considered as restricting the space of the mapping function to reduce its complexity, to allow ML models to generalize well for new cases.

Many materials databases have been exploited using the regression approach to develop user-friendly ML models for a plethora of materials properties. The list is truly unending; examples range from basic thermodynamic (formation or free energies), electrical (dielectric strength[52], electrical conductivity[53], electric breakdown[54,55]), mechanical (tensile strengths[56,57], elastic constants[58], fracture toughness[59]), thermal (heat capacity[60], vibrational spectra and entropy[61], thermal conductivity[62]) properties to more exotic or application-oriented cases, such as superconductivity[63], ferroelectric Curie temperatures[64], properties of topological insulators[65], thermoelectric figure of merit[66] and hopping barriers in Li-ion batteries[67]. Similar to the strategies adopted in drug discovery in the past, all regression studies have a common goal: exploit the existing materials databases for ML model development, quickly estimate properties of hundreds to thousands of new candidate materials using the developed ML model and virtually screen promising candidates for experimental validation. This has lead to accelerated materials discovery, with notable success stories of superhard ceramics[68], thermoelectrics[66], interfaces with high thermal resistance[69], high-strength high-entropy alloys[26] and metallic glasses[70].

Another good example of using ML for virtual screening is the Polymer Genome suite, which uses a hierarchy of intermediate-level fingerprints (such as atomic fragments and morphological descriptors) and the GPR algorithm to learn models for over 20 important polymer properties, including dielectric constant, band gap and glass-transition temperature[71]. The model accuracy varies depending on the quality, source and quantity of the property datasets. Model predictions based on systematically computed data tend to have low prediction errors. For instance, correlation coefficients R^2 of 0.98, 0.97 and almost 1 are obtained for band gap, electron affinity and atomization energy prediction models, respectively, trained on data arising from computations. By contrast, models based on experimental datasets can incorporate additional noise owing to varying empirical conditions, causing relatively high prediction errors, with R^2 values of 0.97, 0.94 and 0.89 for glass-transition temperature, melting temperature and dielectric constant, respectively. All models, however, have reasonable accuracy that allows virtual screening of thousands to millions of polymer candidates in an efficient manner. For example, ML has been used to screen polymers for high-energy-density-capacitor applications, that is, polymers with large band gap and dielectric constant, with selected cases validated via direct experimental synthesis[24,72].

In some studies, the different aspects of the ML-based materials-discovery process — the materials database generation, regression model development, candidate screening and experimental validation — are all performed collectively in an iterative manner using active learning. Given that materials discovery is inherently 'extrapolative', such active-learning approaches are extremely successful[64,73]. A few other ML techniques (such as leave-one-cluster-out cross-validation[74]) that focus on better generalizations of the surrogate models are also adept for such problems. Beyond materials discovery, regression approaches have also helped accelerate experiments and quantum-mechanical calculations, as discussed in later sections.

The materials fingerprints can also be used to quantify the similarity between different materials. Common distance metrics include the Euclidean, Manhattan and Tanimoto coefficients[75]. The choice of the distance metric depends on the problem, but the Tanimoto and cosine coefficients have been found to perform best for similarity calculations. The Euclidean and Manhattan distances are not recommended, although their variability and diversity from other distance metrics might be advantageous for data fusion[75].

Information regarding a materials property of interest can exist at several levels of fidelities, with the measurements varying in terms of their cost and accuracy. For example, structural parameters can be either estimated theoretically (quick but low fidelity) or measured using diffraction experiments (time-consuming but high fidelity) after careful synthesis. For such problems, multifidelity information-fusion approaches may be adopted, wherein information available at different levels of fidelity is combined to build a powerful model that makes predictions at the highest level of fidelity. In this approach, the traditional regression problem is modified to build two complementary models, with the first modelling the low-fidelity data and the other learning the difference between the high-fidelity and the low-fidelity data. The prediction for a new case is given by the sum of the low-fidelity model scaled by a factor and the difference model. Not only can this approach be extended to more than two levels of fidelity but it is a powerful way to combine large materials databases of low-fidelity theoretical property estimates with small-sized but accurate empirical measurements. This approach was used for the problem of predicting the tendency of a polymer to crystallize[76]. ML models were obtained by combining large amounts of low-fidelity estimates from theoretical models with small amounts of high-fidelity crystallinity values obtained from expensive X-ray diffraction measurements. The multifidelity model achieved a root-mean-square error of 12.58% in crystallinity prediction, lower than the value of 17.04% for the single-fidelity GPR model. Other studies on learning band-gap and dopant-formation energies using information from multiple levels of exchange-correlation functional approximations within the density functional theory (DFT) formulation are also noted[77–80].

LASSO, SISSO and symbolic regression. In materials science, we are often concerned with establishing general and simple structure–property relations (such as the Hall–Petch relation for grain-boundary strengthening, the Hume–Rothery rules for solid solutions and the Goldschmidt tolerance factor for the stability of perovskites), with the aim of finding easily interpretable and understandable models. However, the common regression techniques discussed earlier, owing to their high complexity and black-box nature, offer meagre model interpretability. Consequently, the relevance of the optimized mapping function $\hat{f}\,(\,.\,)$ and material fingerprints towards accurate property predictions is difficult to decipher, and mostly ignored in practice. Thus, for problems where one is interested in finding a simple, yet accurate, ML model that can be understood (or justified) in terms of known materials or physical principles, methods such as the least absolute shrinkage and selection operator (LASSO), sure independence screening and sparsifying operator (SISSO) and symbolic regression are more appropriate. Furthermore, the ensuing model interpretability can provide hints regarding the scenarios in which the model is expected to be accurate, and those in which it can fail.

Derived from the field of compressed sensing, LASSO[81] can be used to determine an explicit functional form of a materials property in terms of easily accessible fingerprints (or descriptors) or help to find those fingerprints that most strongly influence the target materials property. In this technique, an enormous list of material features (~10^5) is generated using a few relevant chemical descriptors (referred to as the primary features) and compounding them through various mathematical transformations (such as x^2, log (x) and exp$x$) and operations (such as addition and multiplication), as shown in FIG. 4c. Given this set of extremely high-dimensional (~10^5) fingerprint vector **F** and corresponding target property values **P**, LASSO aims to find a substantially small subset of features (say, 2 to 5) that best describe the data. It solves this non-deterministic polynomial-time NP-hard problem by recasting it into a convex minimization problem: $\arg\min_c (\|\mathbf{P} - \mathbf{F}c\|_2^2 + \lambda\|c\|_s)$, wherein $\|c\|_s$ indicates the Manhattan distance of all non-zero coefficients c and λ is the regularization parameter. The second term in the above expression differentiates LASSO from ridge regression and drives the coefficients of many features to be exactly zero, thereby, resulting in feature selection. Further, the larger the λ parameter, the lower the dimensionality of the optimal solution. In the end, an analytical model that is a weighted linear sum of the small subset of selected features is obtained, which can be easily interpreted, since it is based on simple mathematical transformations to the primary features. New chemical and physical insights can also be extracted. For instance, LASSO was used[55] to learn the electrical breakdown field (F_b) in materials and found it to be neatly described by just two key descriptors, band gap (E_g) and phonon cut-off frequency (ω_{max}), using the expression $F_b = 24.442\exp(0.315E_g\omega_{max})$ MV m^{-1}, with an R^2 coefficient of 0.69 on a dataset with the electrical breakdown field value spanning four orders of magnitude. Similarly, LASSO has been utilized to find simple physical descriptors that determine whether a given binary octet crystallizes in the rocksalt or zincblende structure[82].

The LASSO approach, however, breaks down when the space of candidate features gets very large (>10^8) and/or when features are correlated. To overcome this issue, sure independence screening (SIS) was introduced[83] to iteratively select a reasonably sized subspace of features (up to 10^5) using their correlation with the target property (or residual errors obtained from a model based on a currently selected set of feature candidates). Once the original feature space is reduced, the SIS operation is followed by a sparsifying operation (SO), just like LASSO, resulting, overall, in the SISSO approach. The main advantage of SISSO over LASSO is the gain in model accuracy, owing to the larger feature space explored. Thus, SISSO has been successfully used for the discovery of new relations that differentiate metals from insulators[83] or identifying single or double perovskites with low formation energies[84]. Particularly, in the latter example, the researchers introduced a new definition of tolerance factor ($\tau = \frac{r_X}{r_B} - n_A\left(n_A - \frac{r_A/r_B}{\ln(r_A/r_B)}\right)$) that predicts the stability of the perovskite structure with 92% accuracy, as compared with 72% achieved with the well-known Goldschmidt tolerance factor[85]; here, n_A is the oxidation state of A and r_i the ionic radius of ion i in the compound ABX_3. Interestingly, the new tolerance factor was found to be equally applicable to double perovskites ($A_2BB'X_6$), although it was identified using data on single perovskites (ABX_3), demonstrating the generality of this approach. More importantly, both LASSO and SISSO approaches have been used to rationally distil key materials features that are most correlated with target properties, providing simple and practical guidelines for experiments[86,87].

Another direct method of extracting an explicit functional form of materials properties from the available data is symbolic regression[88,89]. In contrast to LASSO and SISSO, which perform a rather controlled search on a predefined list of candidate features, symbolic regression completes an unconstrained search on the function space spanned by a collection of given function building blocks, such as mathematical operators, analytic functions, material-property variables and constants, to find the most appropriate solution (with minimum training error). The search is performed based on evolutionary methods, such as genetic programming or its more advanced variants, including grammar-guided genetic programming, grammatical evolution and Cartesian genetic programming. An important aspect of symbolic regression is the representation of a (complex) function in the tree structure, with mathematical operators (+, −, ×, ÷) occupying the non-terminal nodes (branches), and variables and constants forming the terminal nodes (leaves). Starting from a generation of candidate solution functions, all represented in their respective tree structures, the usual evolutionary operations of crossover (pruning and mixing between a generation of candidates), mutation (random alterations within a candidate) and selection (retention of candidates with low errors on the training set) are applied iteratively to reach an optimal solution. Here, the mutation and crossover operations are performed, respectively, by random substitutions in part of the candidate tree structure or by replacing subtrees from another solution candidate. The efficiency of symbolic regression can be drastically improved using two strategies: first, the crossover and mutation operations can be constrained to generate only physically meaningful candidate functions (for example, with correct dimensional units) using predefined rules in the form of context-free grammar, and, second, prior domain knowledge can be incorporated by rationally choosing the initial set (first generation) of solution candidates (or specific analytical functional forms). Symbolic regression has been used to discover functional forms of interatomic potentials[90], phenomenological models describing relationships between different materials properties (for example, flow stress and temperature-dependent strain rate[91], or mechanical strength and creep[92]). Overall, the ability of this method to efficiently search for simple and interpretable mathematical models resonates well with the scientific approach of formulating principles and theories using concise equations.

Deep learning. Although many materials databases are relatively small (<10,000) and ML algorithms provide good predictive performance, certain materials problems are best solved using deep learning methods[1]. These include problems that are highly data-intensive (involving millions of data points, for example, resolving diffraction data, image segmentation and learning quantum-mechanical properties, such as electronic charge density, band structure or atomic forces), comprise a large number of fingerprint dimensions or are simply outside the scope of what common ML algorithms can achieve. The superiority of deep learning models is evident through their success in the field of computer vision, AI-played games and speech recognition, and, more appositely, for this Review, in drug or polymer discovery, development of interatomic potentials, mining materials literature using NLP and solving the inverse problem of materials design using generative models. Further, some of the best-performing property-prediction models are based on deep learning. This can be attributed to their highly flexible nature, which is also reflected in a variety of available model architectures, each modified to tackle a specific type of input-data structure. We emphasize that, as more materials knowledge becomes available in a structured manner, such as online repositories and databases, deep learning models are poised to make a great impact in materials science, especially in areas that inherently involve large amounts of data. Below, we briefly describe some common types of deep learning methods that have been particularly useful from a materials viewpoint.

Deep learning uses multiple layers of simple but nonlinear modules to transform raw data (such as element type or atomic positions) into a more abstract representation that can be used to learn complex functions (such as the potential-energy surface). Fully connected NNs are the most basic type of deep learning models, wherein multiple layers of neurons are combined to transform the data, with each neuron being a weighted combination of its respective inputs, followed by a nonlinear activation unit (FIG. 4d). In modern applications of NNs, the number of layers of a model can be very large to achieve a better representation power, hence, the name deep learning. Unfortunately, the higher representation power is also accompanied by a need for larger training datasets to attain good accuracy. The NN weights are learned by optimizing a loss function measuring the difference between the NN outputs and the desired targets. NNs can also be viewed as a model with nonlinear basis functions in which the basis functions are themselves learned from the data. By contrast, for other nonlinear models, such as kernel methods, the nonlinear basis functions are typically fixed beforehand and only weighted combinations of these basis functions are learned from the data. In materials science, NNs have been particularly successful for simulation acceleration, irrespective of the type of modelling technique (quantum-mechanical DFT[4], classical molecular dynamics[93], coarse-grained or finite-element modelling[94]). This is mainly due to the data-intensive nature of these problems, with the added capability of generating new data on demand by running actual simulations based on physical models. Thus, great progress has been achieved in building interatomic potentials, force fields and even learning density functionals or electronic structure using NNs. For example, using only the Pauling electronegativities and Shannon ionic radii of the constituting elemental species as inputs to simple architectures (one or two hidden layers) of NNs, it was possible to learn the DFT formation energies of $C_3A_2D_3O_{12}$ garnets and ABO_3 perovskites with a mean absolute error (MAE) of 7–10 meV per atom and 20–34 meV per atom, respectively[95]. The achieved accuracy is far superior to that of other ML models (~100 meV per atom) and is close to the errors (~24 meV per atom) in the DFT-computed formation energies of ternary oxides relative to the experiments. On similar lines, a NN (termed ElemNet) was trained to learn DFT-computed formation energies of 275,759 compounds from the Open Quantum Materials Database (OQMD) using only their elemental compositions as input[96]. The NN model not only achieved better prediction accuracy but also predicted phase diagrams of new chemical systems more precisely than conventional ML models based on manually constructed features leveraging physics and domain knowledge. Beyond property prediction, deep NNs have made possible the autonomous design and synthesis of molecular systems, accelerated high-throughput experiments and allowed real-time phase retrieval from diffraction data, as discussed in later sections.

Inherent structure in materials science problems can be exploited to reduce the amount of data needed to fit deep learning models. Examples are, for images, translation-variant features in the form of spatial local filters; for sequences, such as SMILES (simplified molecular-input line-entry system) strings, features respecting the sequential nature of the data; or, for molecular materials graphs, features aggregating local graph patterns. Deep learning models, such as convolution neural networks (CNNs), recurrent neural networks (RNNs) and graph neural networks (GNNs, FIG. 4d), can be designed to specifically take these input data structures into account. Furthermore, tasks such as uncertainty estimation and generation of samples other than classifications and regressions may be needed in a materials science problem; in such cases, deep learning models can be combined with probabilistic graphical model methods, using variational autoencoders (VAEs) and generative adversarial networks (GANs).

More specifically, CNNs are NNs designed to learn features from data with grid-like topology (such as images, diffraction patterns, microstructures or even molecular or polymer SMILES). This goal is achieved by applying layers of a mathematical operation called convolution, which involves sliding through a parametric kernel throughout the input data, evaluating their dot product and applying a nonlinear activation function.

As with basic NNs, the weights of the convolution kernel are learned during the model training, allowing extraction of important features from the input data. Notably, CNNs are proficient at extracting translationally invariant features; for instance, they can detect phase boundaries irrespective of their location and orientation within the input microstructure image. CNNs were among the first deep networks that surpassed traditional ML methods, achieving state-of-the-art performance in many computer vision problems, including object detection and recognition, and have been deployed for commercial applications. The reverberation of the impact of CNNs is also felt in materials science, especially in the context of image segmentation[97], fault and failure detection[98], learning microstructure–property relations[99] and developing property-prediction models[100], among other examples[101]. CNNs were also applied to full-Heusler compounds, using only the position of the constituting elements in the periodic table, and could simultaneously learn lattice parameters and the enthalpy of formation with a MAE of 7 meV per atom, which is within the precision of the DFT-computed training data[102].

RNNs are a class of NNs designed for processing sequential data, such as natural languages, time series or even molecular and protein structures. An RNN repeatedly applies a NN module to each token of the input sequence, extracting important signals from the sequential data token by token. RNNs and their variants, such as long short-term memory, have recently led to state-of-the-art performance in many sequence-modelling problems in the areas of speech recognition, NLP and time series. In materials science, they have been most useful for extracting materials databases as part of the NLP pipeline. Further, they are being increasingly used in tandem with generative models (discussed below) for molecular and polymer discovery by directly predicting sequences of molecular connections or SMILES with target properties of interest.

GNNs are NNs designed for processing graph data, involving nodes and edges, each with their own set of attributes. Molecules and materials can also be intuitively represented as graphs, with atoms forming the nodes and bonds the edges of the graph. GNNs take such a graph as input and iteratively perform nonlinear message-passing operations parameterized as NNs, wherein attribute information of nodes and edges are mixed with those of their neighbours, to learn essential features of the graphs. GNNs have been primarily used to learn computational data, such as formation energies and band gaps of materials or molecules[103], predict synthesis pathways using possible chemical reactions[104,105] and 'generate' hypothetical molecules with superior properties[106], and are among the top performers for many benchmark materials datasets. For instance, CNNs and pooling layers were used on graph representations of crystals to automatically extract optimum material representations, which were mapped to different properties (including formation energy, band gap, Fermi energy, bulk and shear moduli) using fully connected NN layers[107]. A remarkably low MAE of 39 meV per atom for the formation energy using a dataset of 28,046 entries from the Materials Project was achieved. Building on this idea, a generalized graph-based approach was presented for both molecules and crystals, capable of even incorporating global variables, such as pressure and temperature, beyond the local information on atoms and bonds[108]. The approach was based on a series of NN-based update functions (termed 'MEGNet' blocks) that allow information on different nodes (atoms), edges (bonds) and global variables to mix, with the final readout operation reducing the output graph to a scalar or vector target property. Some of the best reported predictions on the QM9 molecular dataset and Materials Project crystal dataset have been achieved using this model with, for example, a MAE of 10 meV per atom and 28 meV per atom for the Gibbs free energy and formation energy, respectively. Besides small molecular and crystal graphs, GNNs can also be used to analyse much larger networks, such as online social networks and knowledge graphs, and tailored to perform graph node classification and edge prediction. In another interesting work[109], a materials stability graph was constructed by considering different materials as nodes and their convex-hull thermodynamic free energy tie lines as edges using data from the OQMD. By analysing the time evolution of such a stability graph, that is, by successively adding information on newly synthesized materials, the synthesizability likelihood of computer-generated hypothetical materials could be predicted.

Another important application of GNNs is to directly learn the material fingerprint, especially in the context of the atomistic problems that take the material structure as an input. As discussed earlier, materials datasets should be first fingerprinted into fixed-dimensional vectors to be ready for the application of ML methods. Mostly, such transformations are done by human experts without taking into account the downstream optimization problem, and involve high-dimensional representations to achieve sufficient accuracy. By contrast, material graphs can be passed as an input to GNNs, where network-based, nonlinear message-passing operations are performed to learn their material fingerprints. The details of the message-passing operator define the type of GNN, encoding the nature of the prior knowledge into the network, and allowing it to better learn certain types of data. For instance, gated graph NNs use a learned weight (or gate) to aggregate messages from neighbours, and graph convolution NNs use a graph Laplacian in the message aggregation[110]. Although GNNs have already demonstrated their mettle for several materials science problems, their use is expected grow considerably, owing to their versatility to learn different types of materials fingerprints and properties collectively.

VAEs[111] are a type of generative model that learns to encode or decode a collection of inputs to and from a latent space. This is achieved by using deep NNs (feed-forward, convolution or recurrent type) to represent the encoding and the decoding units, with the constraint that the encoder lowers the dimensionality of (or compresses) the input data, while the decoder performs the decompression operation to reconstruct the input sequence. As an outcome of this data compression and expansion exercise, the VAE learns the underlying properties of the data itself.

Akin to VAEs, GANs[112] are generative models consisting of a generator and a discriminator unit. A GAN learns to generate new samples from the underlying data distribution (as supplied using the training data) through a game between the generator and the discriminator units; the generator tries to fool the discriminator by constructing new (fake) examples that are close to the training data, while the discriminator attempts to catch the deceit of the generator by correctly separating the generated samples from the real examples. Typically, both the generator and the discriminator units are parameterized by NNs and learn in a min-max fashion by competing against each other. The way the decoder and the encoder units of the VAE and the generator unit of the GAN solve important materials problems is discussed in the section on emerging developments.

Infrastructure transformation

The laboratory experiences of researchers in the future are expected to be distinctly different from the present, owing to the introduction of several types of ML agents in modern materials science labs. This applies to both computational and experimental efforts. Below, we review a few emerging developments.

ML-assisted acceleration of simulations. Most simulation software — be it electromagnetic simulators, finite-element solvers or simulators based on quantum mechanics (such as DFT) — may be viewed as modular tools that produce materials property outputs based on the input simulation conditions (FIG. 5). For instance, a typical DFT code takes as input the atomic coordinates of a molecular or material system and produces a variety of outputs, such as the electronic wavefunctions, charge density and energy levels (the primary output), total energies, atomic forces and unit cell stresses (the secondary output), and other derived properties, such as formation energies, elastic constants and dielectric constants (the tertiary output).

Thus, treating the functioning of DFT as an input–output problem, an enormous body of effort has emerged to build surrogate models. For instance, the ability to efficiently predict the tertiary outputs of DFT has enabled several examples of accelerated materials discoveries. Likewise, surrogate models of the secondary outputs of DFT have lead to ML force fields that have the potential to overcome several major hurdles encountered by both classical and quantum molecular dynamics simulations. Most recently, initial steps have been taken to learn and predict the primary outputs of DFT using deep NNs. These efforts may soon lead to DFT emulators that can mimic DFT in all respects, while being several orders of magnitude faster. Below, we review important works based on the level (primary, secondary or tertiary) of materials property being targeted.

We have discussed how regression algorithms have been exploited to learn several tertiary properties (such as mechanical, thermal and formation energies) and allow virtual screening of materials with desired properties. Similarly, the application of LASSO, SISSO and symbolic regression methods to tertiary properties have provided new insights into the materials structure–property relations. A common feature across many tertiary property ML models is that they are based on relatively simple kernel or decision tree methods, owing to the small amounts of available property data. By contrast, the secondary and the primary properties models utilize more complex NN methods to learn from the large amounts of available data. Several tertiary property models can also be formulated as classification problems, such as stable versus unstable perovskite structures[84,113] or determining the most likely lattice symmetry of a defect-containing structure[114].

Learning secondary materials properties (atomic forces or energies) have led to a revolution in the development of interatomic potentials and force fields. Since the early 1960s, classical interatomic potentials have been used to study materials processes and behaviours

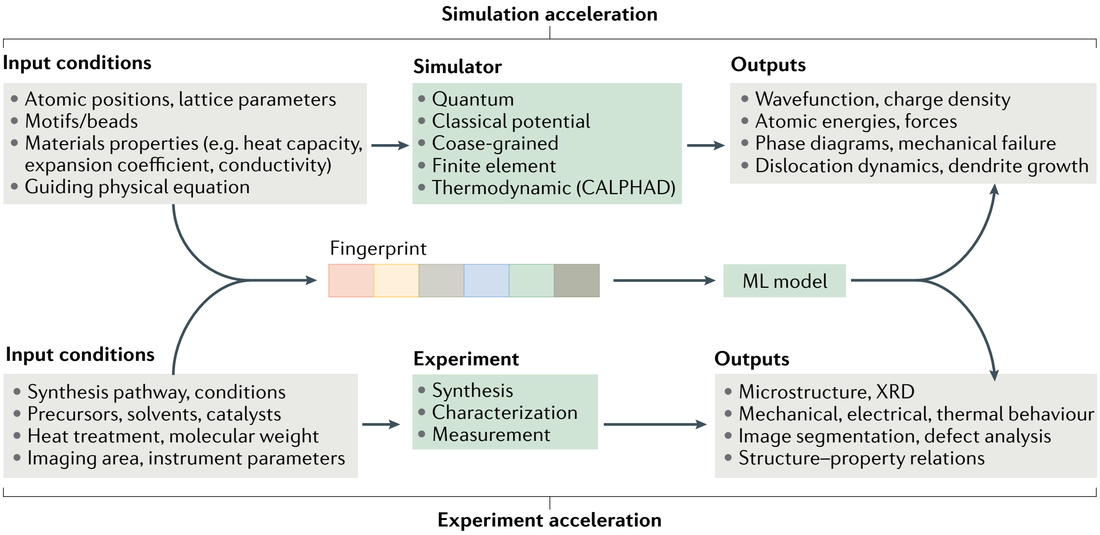

Simulation acceleration
Input conditions
Atomic positions, lattice parameters
Motifs/beads
Materials properties (e.g. heat capacity, expansion coefficient, conductivity)
Guiding physical equation
Simulator
Quantum
Classical potential
Coase-grained
Finite element
Thermodynamic (CALPHAD)
Outputs
Wavefunction, charge density
Atomic energies, forces
Phase diagrams, mechanical failure
Dislocation dynamics, dendrite growth
Fingerprint
ML model
Input conditions
Synthesis pathway, conditions
Precursors, solvents, catalysts
Heat treatment, molecular weight
Imaging area, instrument parameters
Experiment
Synthesis
Characterization
Measurement
Outputs
Microstructure, XRD
Mechanical, electrical, thermal behaviour
Image segmentation, defect analysis
Structure–property relations
Experiment acceleration

Fig. 5 | **Impact of machine learning on the materials research infrastructure.** Machine learning (ML) is accelerating and building efficiencies within several operational and infrastructural aspects of computational and experimental (synthesis and characterization) materials research efforts. XRD, X-ray diffraction.

using molecular dynamics and Monte Carlo simulations. Their efficiency and simplicity easily overshadows other computationally demanding quantum-mechanical schemes (such as DFT), allowing them to study materials properties at time and length scales that can be paralleled in experiments. Historically, these classical potentials have been constructed by fitting an approximate analytical energy expression to reproduce tabulated empirical (lattice parameters, elastic constants) or quantum-mechanical (phase energetics, atomic forces) data. The complexity or functional form of the potential is chosen based on the interactions (ionic, covalent, dispersive etc.) believed to be dominant, with the number of fitting parameters varying from two (Lennard–Jones[115]) to nearly 100 (ReaxFF[116]). Unfortunately, little guidance is available on how to tune these parameters, with developers mostly relying on chemical intuition or traditional heuristics.

However, two general directions have emerged in which ML is pushing the boundaries of this field: the efficient and autonomous parameterization of potentials with predefined functional forms using active learning or evolutionary algorithms, and the direct learning of the functional form of these potentials from the available high-fidelity quantum-mechanical or empirical datasets. For the first point, Bayesian optimization and genetic algorithms have been successfully used to parameterize potentials for difficult systems, such as glassy silica[117], water[118] and WSe_2 (REF.[119]). They have also led to the creation of autonomous workflows that can be used by novice users to develop their own potentials by following through necessary stages of development, such as generating appropriate training datasets, parameter optimization and, finally, cross-validating their potential predictions[120]. This democratizes the development process and reduces their dependency on a handful of developers.

Despite their popularity, potentials with predefined functional form are inherently inflexible, limiting their accuracy and transferability. This brings us to the second point, the direct learning of the potential functional form, that is, the establishment of a direct mapping from the atomic neighbourhood, or fingerprint, to atomic energies — and their sum to total potential energy — using reference quantum-mechanical data. As no prior restriction is imposed, this methodology is quite general and can be used to learn energy functionals of diverse materials (metals, ceramics, alloys, polymers) involving different atomic interactions with minimal human interference. This also overcomes the limitations of traditional potentials designed to capture specific interatomic interactions. As compared with other examples of ML in materials science, this field is quite mature, with tremendous efforts devoted to fingerprint development, effective sampling of training data and physics-informed ML architectures[3,50,121–124]. This approach has been successfully used for numerous systems, including elemental bulk materials (such as Al (REFS[125,126]), C (REFS[127,128]), Li (REF.[129]), Si (REFS[130,131]), Fe (REF.[132]) and Zr (REF.[133])), alloys[134], metallic clusters[135,136], semiconductors[137], oxides[138,139], water[140,141] and organic molecules[142,143], and complex phenomena, such as diffusion[144] and phase equilibrium. Different software packages (such as RuNNer[130], GAP suite[145], AMP[146], AGNI[125], DeepMD[147], AENet[139], MTP[148], SchNetPack[149], N2P2 (REF.[150]) and SNAP[151]) that allow users to train and validate their own potentials for a particular system have also been released. Interestingly, many of the ML-based potentials report low errors, on the order of 10 meV per atom and 0.3 eV Å^{-1} for energies and forces, respectively, at least for structures that are 'close' to the training set[126,152]. The generality of these potentials has been demonstrated for complex, multi-elemental transition-metal oxides and biomolecules containing up to 11 chemical species[153], which is far more than what can be easily targeted using traditional classical potentials.

Despite such a large body of work, there are some basic challenges in this field. First, only limited comparative studies that clearly highlight the accuracy versus speed trade-off of the various fingerprinting and/or ML model architectures have been put forth[152]. Second, it is unclear how one can assess the physical and chemical domain of applicability of ML-based potentials; classical potentials suffer from the same limitation. Bayesian-based uncertainty estimates[154] or estimates based on deviations of energy and force predictions from ensemble methods[155] have shown promise. Active-learning[156] and transfer-learning[157] approaches have also been used to continuously expand the domain of applicability of these potentials. Lastly, there exists no systematic and efficient pathway for the generation of high-fidelity training data, especially for potentials capturing complex interactions involving grain boundaries, surfaces, defects or multiple elements. Standardized datasets allowing fair comparisons across different ML-based potentials should also be developed.

Lastly, in terms of primary properties, ML has been used to learn the electronic charge density, electronic density of states or density functionals themselves, thus, going to the very heart of DFT. A major bottleneck of DFT is the computationally demanding nature of the self-consistent solution to the Kohn–Sham (KS) equations (specifically, having to orthogonalize single-particle eigenvectors), requiring high-performance computers. However, the primary outputs of the KS-DFT, that is, the electronic charge density or one-electron wavefunctions, can be learned using ML methods, bypassing the need to compute expensive self-consistent solutions. Importantly, these efforts are supported by the famous Hohenberg–Kohn theorem that guarantees a unique functional mapping between the structure of a material (in terms of nuclear potential) and its ground-state charge-density distribution, which, in principle, can be learned using ML. The plane-wave basis representation of the charge density was used to map the corresponding basis function coefficients to the nuclear potentials using KRR[158]. Thus, for a new structure, it is sufficient to input the corresponding nuclear potential to obtain the associated charge-density output in terms of the basis coefficients, from which the total energy of the system can also be obtained. Representation of the charge density as a sum of atom-centred basis functions, rather than the plane-wave basis, has been suggested for better transferability of the method to large and flexible systems[159]. In another approach, the charge density and local density of states (DOS) predictions at a grid point were obtained by mapping them directly to the atomic neighbourhood using deep NNs and RNNs, respectively[160]. Summing up the local DOS overall grid points resulted in the total DOS, which, when combined with charge-density predictions, can be utilized to directly obtain the total energy of the system, eliminating the need to solve KS-DFT equations. Advantages due to the strictly linear scaling of the ML approach, as compared with the quadratic scaling (at best) of DFT, were also demonstrated for the case of Al and polyethylene slabs[160]. Moreover, the generalizability of this approach was recently demonstrated for over 70 different hydrocarbons, including alkanes, alkynes, cyclo-groups and even polymers[161].

ML is also assisting in solving the most fundamental question of DFT, that is, the functional form of the exchange-correlation (xc) functional. Using the auto-differentiation functionality, NNs have been trained to reproduce not only the correct xc energy but also the corresponding xc potential, obtained as functional derivatives with respect to the density[162]. Similarly, CNNs have been used as surrogate xc functionals that take electron density as input, extract the necessary electronic features and map them to the desired xc energy density[163]. Other attempts to learn the kinetic energy functionals and provide an alternative orbital-free DFT solution have also been made[164].

In principle, the primary property predictions from a highly accurate ML model can be plugged into relevant physical equations to derive secondary and tertiary properties. But, in most studies, the ML models did not translate well to other properties. Perhaps a unified model approach using multitask NNs or multi-fidelity co-kriging and learning all the different levels of properties together could be more successful.

ML-assisted enhancements to laboratories. Materials scientists have absorbed different AI techniques in their labs to support activities ranging from experiments guided by combinatorial searches and automated high-throughput experimentation, to efficient acquisition and (real-time) analysis of instrument data. As discussed, the foremost change has been of a psychological nature: materials science has moved on from a trial-and-error approach to materials discovery to more informed combinatorial searches, wherein either computational (DFT) or ML-based screening is done first to identify the most promising material candidates for synthesis and design in the labs.

ML methods have also been integrated into several instrumentation facilities to either accelerate the experiment itself or allow for on-the-fly data analysis. For instance, a key practice in materials science is to find relations between the microstructural features of a material (such as particle sizes and shapes and grain-boundary pinning points) and its macroscopic properties (such as yield strength and corrosion resistance). As advances in materials-characterization technologies, including microscopy, spectroscopy and macroscopic testing, have led to a proliferation of materials imaging data, algorithmic tools that quickly process these images and extract relevant materials features are necessary. Common microstructure-segmentation methods rely on either specialized image-processing pipelines requiring expert parameter tuning or are evaluated manually and subjectively. Accordingly, several ML strategies based on Bayesian inference (with maximum a posteriori, maximizer of the posterior marginals or minimum mean-squared-error criteria) have been adopted for automated image segmentation[165,166]. Deep learning has also shown promise in this area, with the ability to identify constituent phases of complex microstructures. A novel CNN-based method, PixelNet[97], could successfully segment steel micrographs into regions of grain-boundary carbide, spheroidized particle matrix, particle-free grain-boundary denuded zone and Widmanstätten cementite, and then segment cementite particles within the spheroidized particle matrix. Because segmentation is a pixel-level task, PixelNet was trained to produce a latent representation of each pixel, instead of the entire image. This pixel latent representation was mapped to the corresponding pixel-level target using a simple deep neural network. Although the authors noted some difference in the model predictions and the human-annotated micrographs, especially for particles with radii smaller than five pixels, this study demonstrates that AI-driven microstructure-segmentation systems can be combined with other emerging automated microscopy capabilities for high-throughput investigations of microstructure-based materials design and optimization. In more industrial settings, automated image analysis has also led to easy identification of defects, cracks or corrosion in structural materials, such as railway tracks and asphalt pavements[167].

Statistical measures (n-point correlation functions, entropic descriptors) serving as descriptors for microstructural images (obtained from either experiments or phase-field models) have been combined with dimensionality-reduction techniques such as principal component analysis to establish processing–structure–property–performance (PSPP) relations[168,169]. Deep learning methods have been used to learn both the homogenization — information transfer from smaller length scales (structural information) to bigger length scales (macroscopic properties) — and the localization — information transfer from bigger length scales (macroscopic stress) to smaller length scales (load distribution across structure) — aspects of PSPP[170,171]. For example, a CNN was used to map microstructural images of two-phase composites to their macroscopic stiffness with the training data obtained from micromechanical finite-element simulations. Interestingly, the accuracy of the deep learning model surpassed that of commonly employed physics-based approaches or rule of mixtures method, and other physics-inspired data science approaches with handcrafted microstructural fingerprints, clearly demonstrating that deep learning could be a powerful tool for representation learning in materials science. The problem of materials design can also be translated as that of microstructural design, where the aim is to find an optimal microstructure that leads to the desired macroscopic properties. To this end, GANs[172] have been trained to represent microstructures in a low-dimensional latent space, which was searched using Bayesian optimization to find the latent vector (and the corresponding GAN-generated microstructure) that resulted in the desired optical absorption performance.

ML-based methods are also being used to accelerate materials imaging. For example, CNNs have been shown to surpass traditional methods for indexing electron backscatter diffraction patterns to determine crystal orientations. Not only were CNNs found to be more accurate but they were also more robust and cheaper than traditional indexing methods that are either susceptible to noise (Hough transformation method) or extremely computationally demanding (dictionary-based indexing)[173]. In another study, a substantial acceleration in the measurement time of a scanning probe microscope (SPM) was achieved by using Bayesian uncertainty to determine the next measurement point, rather than relying on traditional measurements made over an exhaustive spatial grid[174]. Using only ~30% of the original data, high-quality reconstruction was achieved. This study set the stage for the use of Gaussian processes for the development of automated experiments, where an appropriate definition of the acquisition function can be used to drive the measurements. The concept of AI-driven SPM was recently realized for a scanning tunnelling microscope that used a CNN classifier to assess the quality of the acquired images as 'good' or 'bad', a deep reinforcement learning agent to reliably condition the state of the probe such that the acquired images were deemed as good by the classifier and an algorithmic approach to sample different regions of the material. Only images that were classified as good were processed and saved to the hard disk. This approach paves the way for advanced imaging methods hardly feasible by human operation, such as large dataset acquisition spanning multiple days and SPM-based nanolithography.

Other notable examples in this area include ML methods for automated image analysis for the detection and tracking of defects[175] or mesoscopic phase evolution in 2D materials[176], recovery of 3D atomic distortions in a variety of oxide perovskite materials[177] from 2D scanning transmission electron microscope micrographs, solution of the inverse phase problem (that is, recovering the phase information from the measured diffraction intensity in the reciprocal space) in electron microscopy and scattering experiments[178,179], sparse dynamic sampling to reduce image acquisition time[180], prediction of battery cycle life from early-cycle discharge voltage curves[181,182] and on-the-fly analysis of X-ray diffraction patterns for rapid phase identification and distribution analysis[183–185]. Particularly for the last problem of rapid phase identification, efforts have been undertaken to overcome the challenges of peak shifting (due to alloying, or interstitial or substitutional solutions), establishing compositional connectivity, matching with realistic spectra and enforcing physical constraints, such as the Gibbs phase rule[186–188].

Robotics and active learning. The confluence of robotics, computer vision, materials synthesis and characterization, and ML is causing a revolution in materials science. The concept of autonomous or self-functioning materials labs, similar to the idea of driverless cars, chatbots, unmanned aerial vehicles or drones, is now a reality. These autonomous research laboratories (or meta-labs) use AI-driven robotic platforms to synthesize, characterize and measure materials properties, which are then analysed using ML-based surrogate models to optimize parameters for the next experiment[189]. Remarkably, all of these steps are performed in an autonomous manner, with the only input required from the human being the target material of interest. Such autonomous labs offer numerous advantages over traditional materials facilities that rely on manual operation of experimental apparatus: time and cost benefits, experimentation consistency, long hours of operation, and efficient and robust search of the experimental parameter space to achieve optimal materials. It is important that automated labs are not confused with autonomous labs, as, in the former, only predetermined repetitive tasks with well-defined responses are performed, without any ability to hypothesize and conduct new experiments, whereas in the latter, new experiments are autonomously planned and executed, and the resulting information is digested to optimize the next set of experiments.

Although autonomous labs are still in their infancy, there are already notable examples from the fields of biology, pharmacology, chemistry and materials science. The first robotic researcher, Adam, was designed for functional genomic research, measuring growth curves of microbial strains in different environments[190]. In the field of drug discovery, robot scientist Eve was used to intelligently screen drugs for tropical diseases in a high-throughput and economical manner[191]. However, more relevant to materials science is the Autonomous Research System (ARES)[192] that learned to grow single-walled carbon nanotubes at targeted growth rates. The active-learning approach (based on random forest and evolutionary algorithms) is coupled with an automated growth reactor and in situ characterization (Raman spectroscopy) to perform robotically controlled scans of the experimental parameter space (temperature, pressure and gas composition). At each iteration, the ARES surrogate model predicts the experimental growth conditions that can achieve the target growth rate supplied by the user. The system then conducts the experiment at the proposed conditions and measures if the output growth rate matches the target. If any discrepancy is observed, the existing database is updated with the measured rates and the ML models are refined for the next iteration. Experiments are automatically stopped once the measured growth rate matches the target (within predefined standard deviations). In a similar approach, GPR-based models were used to autonomously produce high-quality Bose–Einstein condensates by optimizing the evaporation ramp and using the images of cold atoms as feedback[193]. In another study, a human and robot researcher tasked with the synthesis, crystallization and characterization of a new polyoxometalate compound were compared. The active-learning-based robot researcher outperformed the human researcher in both the explored crystallization space (the experimental conditions that resulted in successful crystallization of the cluster) and

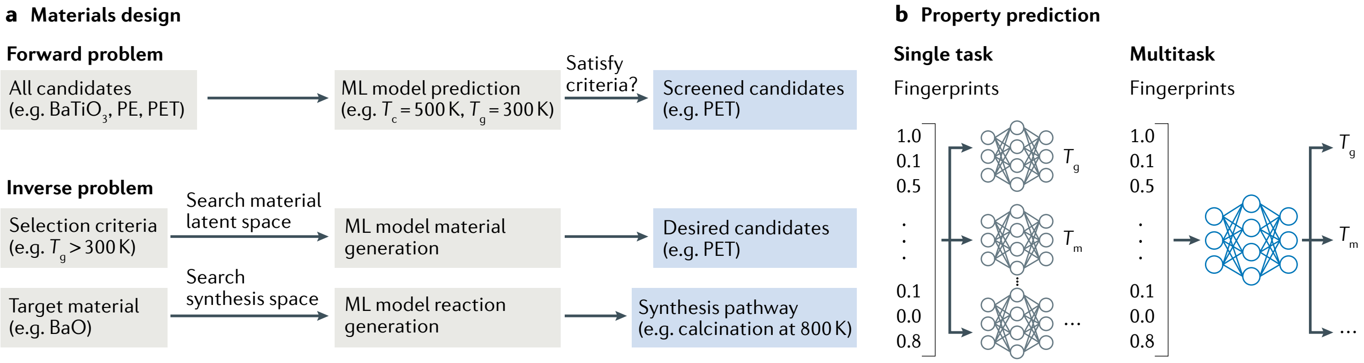

a Materials design
Forward problem
All candidates (e.g. BaTiO$_3$, PE, PET)
ML model prediction (e.g. T_c = 500 K, T_g = 300 K)
Satisfy criteria?
Screened candidates (e.g. PET)
Inverse problem
Selection criteria (e.g. T_g > 300 K)
Search material latent space
ML model material generation
Desired candidates (e.g. PET)
Target material (e.g. BaO)
Search synthesis space
ML model reaction generation
Synthesis pathway (e.g. calcination at 800 K)
b Property prediction
Single task
Fingerprints
1.0
0.1
0.5
0.1
0.0
0.8
T_g
T_m
...
Multitask
Fingerprints
1.0
0.1
0.5
0.1
0.0
0.8
T_g
T_m
...

Fig. 6 | **Opportunities for materials design using advanced machine learning algorithms. a** | The forward and inverse problems are two distinct screening pathways for material selection using machine learning (ML) models; solving the forward problem relies on screening a predefined list of candidates based on their ML property predictions, whereas the inverse problem is tackled by directly generating material candidates starting from the desired property objectives. Synthesis planning and retrosynthesis design can also be viewed as inverse problems with the aim to find the necessary reactants and processing conditions. **b** | The single-task versus the multitask learning approach; whereas the former learns the different materials properties independently, the latter exploits various property correlations to learn a joint and more powerful ML model. PE, polyethylene; PET, polyethylene terephthalate; T_c, Curie temperature; T_g, glass-transition temperature; T_m, melting temperature.

crystallization prediction accuracy, demonstrating the efficiency and superiority of autonomous labs[194]. Other successful applications include autonomous small-angle and grazing-incidence small-angle X-ray scattering experiments[195], autonomous assembly of 2D crystals such as a 29-layer superlattice alternating graphene and boron nitride[196], synthesis and characterization of perovskites with wide band gap[197] and a very general robotically controlled experimental platform capable of synthesizing complex organic molecules[5,198]. Software packages that allow the easy integration of different pieces of autonomous laboratories, such as instrumentations, databases and ML algorithms, have been put forth[199].

Likewise, on the computational side, as the above-mentioned ML schemes mature for different levels of theoretical outputs, autonomous computational labs will emerge[200]. In the future, simulations will be performed in the inverse mode using active learning; that is, researchers will specify the desired property and the chemical space to be explored, and active learning will iteratively perform simulations to produce the requested materials, leading to an autonomous computational lab. Further, for several computationally expensive simulators (such as DFT), ML-based emulators will become available that will output all the important primary, secondary or tertiary outputs in a reliable and quick fashion, without solving the underlying expensive physical equations.

Emerging developments and opportunities

So far, materials scientists have mostly adapted existing ML tools to analyse and solve problems in their field. However, with the community realizing the benefits of the data-oriented approach, and having trained a generation of researchers with experience in ML methods, many innovative ML methods targeted to solve problems unique to materials science are emerging. Here, we review some of these efforts.

Solving inverse problems. The inverse design problem in materials science consists of determining the material compositions, structures or processing conditions that result in a desired property[6] (FIG. 6a). This is generally solved by building a 'forward' model that predicts the property of a given input material, and using this model with an efficient search strategy in the materials space to screen those candidates that are predicted to meet the property targets. For example, KRR-based property-prediction models were used to identify polymer candidates that met the desired band gap or dielectric constant objectives from a predefined list of polymers[201]. However, the materials space is huge, making a robust search using the forward model restricted and impractical. For polymers, a more effective approach consists in combining the forward property-prediction models with combinatorial optimization algorithms, such as evolutionary or genetic algorithms, to directly search the materials space. This is achieved by adding, removing or shuffling the chemical building blocks (such as CH_2, C_6H_6 and COOH) in the polymer monomer to formulate completely new candidates, for which forward model predictions are performed to check if they display promising properties. Several iterations of this strategy help uncover previously unknown polymers, as demonstrated by exemplary work on the design of polymers for high-energy-capacitor applications[72,202] and biodegradable polymers with high glass-transition temperature (T_g)[203].

Another powerful approach is to use VAEs. We mentioned that the encoder and decoder units of VAEs can be respectively used to learn a mapping from discrete material graphs to a continuous vector representation (latent space), and the corresponding back-mapping going from vector representations to the material graphs. Thus, the decoder provides a pathway to solve inverse problem of materials design, directly 'generating' materials that have the desired properties, unlike traditional ML approaches that screen materials that have target properties. Materials generation using the decoder surpasses the constraints of human imagination, precluding the need to predefine a materials search space. Furthermore, the encoder unit of the VAE can be viewed as a fingerprinting function (representation learning), which can be used to drive downstream AI models and optimization methods. A good example is the use of VAEs and GPR to discover polymers with high T_g and band gap for high-energy-capacitor applications[204]. The encoder unit of the VAE was used to fingerprint polymers, which were mapped to their associated T_g or band gap values using GPR during training. For the discovery part, the researchers first searched in the latent space within the vicinity of known polymers with desirable properties, to find latent points predicted by the GPR model to have the required band gap and T_g. Next, they used the decoder part of the VAE to simply obtain the polymer SMILES associated with the selected top latent points to be validated with experimental synthesis and measurements. Similarly, the VAE encoder outputs have been used to guide Bayesian optimization models to search for drug-like molecules[205].

GANs have also been used for solving inverse problems by generating high-dimensional samples, such as natural images[112], molecular graphs[106] or even porous crystalline materials[206]. In the last example, the authors generated over a hundred (new) zeolite structures based on pure silica that were not only structurally reasonable and diverse but also showed high heat of adsorption for methane, which is important, as methane capture is among the most critical applications of porous materials. In the future, VAEs and GANs may be combined with other supervised NNs to enable more advanced and practical semi-supervised learning. Both VAEs and GANs are expected to drastically alter how virtual screening of new materials is performed, as they provide pathways to go beyond a predefined set of candidates and efficiently explore a diverse materials space that can surpass human imagination.

A key challenge in decoding and generating materials from their vector representation is that a valid material has many physical and chemical constraints (for example valency of 2 for O and 4 for C). Thus, the generative model needs to take into account such constraints, which can be incorporated by translating them into semantic and syntactic checks and building syntax-directed VAEs[204,205]. These models use masking operations to force the decoder unit to generate only valid materials. Moving forward, more constraints that incorporate, for example, polymer synthesizability or ML algorithms that go beyond evolutionary methods or VAEs to directly solve the inverse design problem are expected to become widespread.

Synthesis planning and retrosynthesis. Retrosynthesis planning is the procedure of identifying a series of reactions that will lead to the synthesis of a target product. It can be considered as yet another inverse problem: given a material, how do you synthesize it? Formalized by Elias James Corey[207], it is one of the most fundamental problems in organic chemistry and materials science (FIG. 6a). The problem of 'working backwards from the target' is challenging, owing to the size of the search space — the vast numbers of theoretically possible transformations — and, thus, requires skill and creativity. Recently, ML-based retrosynthesis planning algorithms[208] have been designed to tackle this problem with very promising results[209,210].

The simplest formulation of the retrosynthesis problem is to take the target product as input and predict all possible reactants. Essentially, it is the 'reverse' of the reaction prediction problem, wherein, given the reactants (and the associated reagents) as input, the output is the list of possible products. The solution to the reaction prediction problem is obtained through a deductive reasoning process, because the atoms of the product form a subset of reactant atoms. By contrast, retrosynthesis aims to identify the superset of atoms in the target product(s) and, thus, is an abductive-reasoning process, which requires creativity to be solved, making it a harder problem. Thus, recent advances in GNNs for solving the reaction prediction problem[105,211,212] do not transfer to retrosynthesis.

Many computer-aided retrosynthesis design algorithms are completely rule-based systems that suffer from high computational cost and limited coverage of the possible chemical space, or are expert-defined and cannot be translated algorithmically. Despite these limitations, they are very easy to interpret and provide a useful method to encode chemical transformations. To this end, a tool called retrosim[213] has been developed that uses similarities in the fingerprints of molecules and a library of reactions to select the rules to apply. Other approaches rely on classification models (NNs) for this selection task[214]. There have also been recent attempts to directly predict the SMILES representation of the reactants[215,216] (or the products, for the forward reaction prediction problem[217,218]). Albeit simple and expressive, these approaches completely ignore the rich chemistry knowledge developed over the years and rely solely on the huge amount of reaction training data. Further, such models lack interpretable reasoning behind their predictions. Combining the interpretability of the rule-based methods and the scalability and expressiveness of the NNs is the recent approach of conditional graph logic network, in which chemistry knowledge about reaction templates is treated as logic rules and a conditional graphical model is introduced to tolerate some noise in these rules[104].

Although a large amount of reaction data exist for molecules, there is no such well-organized database for materials synthesis, making this problem far more challenging for systems such as metal oxides, porous metal–organic frameworks and polymers. Initial efforts to use NLP pipelines to extract relevant synthesis conditions from literature data, use them to build predictive ML models and then suggest synthesis conditions have shown promise for the design of zeolites with desired density[36], new fabrication routes of previously known[219] and unknown[38] perovskite materials, and formation of titania nanotubes[37]. However, several challenges remain in each stage of the process, including the accurate extraction of synthesis parameters using NLP and the highly sparse and high-dimensional parameter space arising from the large materials chemical and structural diversity. These challenges are further compounded for the case of polymers by inconsistent naming conventions and our lack of knowledge of the polymerization processes.

Physics-informed ML models. ML-based models purely built on data are agnostic of physical and chemical relations. For example, physics requires that the dielectric constant of a material varies inversely with frequency (ignoring resonance effects), the Gibbs free energy of a phase decreases with increasing temperature (keeping other factors, pressure and composition, constant) and the energy of a system is invariant to system translation or rotation. When a ML-based materials model is trained, the hope is that the model will automatically learn such physical and chemical rules implicit in the training data. However, this is generally not true, especially in the 'extrapolative' regime, where the surrogate models have been found to violate these rules terribly. An alternative approach is to constrain (similar to the idea of constraint programming[220]) the space of functions considered by a ML model (during the training), such that only cases that respect the known physical and chemical relations are allowed. In other words, this approach transforms a ML model from a mathematical data-dependent construction to a physics-informed surrogate model. An added advantage of physics-informed ML models is that they require fewer training data, which is important, given that materials data tend to be small but rich in information.

Although it is non-trivial to design ML models that are physically aware, with the procedure strongly dependent on the underlying materials problem, a few notable examples have been put forth. Within the context of developing generalizable interatomic potentials, the NN architecture has been modified to obtain physically meaningful energy trends. This is achieved by forcing the last layer of the NN to have a general physics-inspired form (such as an analytical bond-order potential). Thus, in analogy with how the last sigmoid layer of a NN classifier constrains its output to class probabilities, the last physics-based layer ensures that the output energy and force predictions follow a physically meaningful trend, as dictated by its functional form[221]. Thus, when equation-of-state predictions were made in extrapolative compressive or tensile environments, the physics-based model predictions followed correct energy trends, whereas the purely data-driven model failed miserably. Likewise, other innovative material fingerprints and model architectures have also been proposed that account for invariance and symmetries present in the target property of interest (system energy, atomic forces)[222–225]. For instance, rotationally invariant fingerprints were designed to learn system energies, and a gradient-based ML approach was developed to construct energy-conserving force fields[226].

Symbolic regression has also been used to develop interatomic potentials to search within a physically meaningful hypothesis space, consisting of mathematical operations and functions obtained from potentials developed over the past several decades. This ensures that the learned potential is not only reasonable and interpretable but can also be compared with other manually constructed and well-studied potentials, such as the Lennard-Jones potential for the case of Al (REF.[90]). Another powerful approach to constrain the model space consists in using context-free grammar, which, essentially, describes a set of production rules that the data of a formal language must follow. For example, in the arithmetic expression $a * b$, a should be a number and not a mathematical operator, or for the molecular SMILES c1ccccc?, the symbol ? must be equal to 1 to complete the ring and represent a valid benzene molecule. Transforming the original data into their equivalent context-free-grammar representation and then using it to build the ML models ensures their syntactic validity. Although this approach has been quite successful for generating syntactically valid molecules and polymers, and for restricting the functional space explored in symbolic regression, more efforts are needed to extend it to a large class of materials problems.

Multitask learning and transfer learning. Materials science focuses on multiple properties of the same material: for a new material to be successfully used in an application, it has to meet several property objectives. However, most examples of ML in materials science use separate models for each task (or property), leading to many model parameters, with little information shared across the different models. Instead, ML methods such as multitask learning can be used to learn a shared vector representation for all tasks, and then use the shared representation to predict different output properties collectively (FIG. 6b). In the case of a NN, this can be achieved by having multiple output layers, one for each property, and a common loss function defined as the weighted sum of errors for each property prediction. The multitask approach is expected to perform better than individual ML models because of two reasons: first, it fuses information about multiple properties to exploit their correlations and learn a more informed model, and, second, it has access to a larger dataset than ML models based on individual properties, as, in practice, only a partial set of properties of a material are known.

Polymer solvent and gas-permeability models have been built with a similar philosophy[227–229]. In these works, the authors used a series of chemical and morphological descriptors to fingerprint the polymer, and a one-hot encoding representation for the solvent or gas; for instance, the vector 001000 represents the case of CO_2 permeability from the total pool of six gases, He, H_2, CO_2, O_2, N_2 and CH_4. Because the ML models were aware of the polymer behaviour for several scenarios (solvents or gases) at once, the prediction accuracy was much higher than that of other theoretical models based on the Hildebrand criteria (for polymer/solvent compatibility) or the Robeson upper bound (for gas permeability). We note that multifidelity learning, as we discussed, also uses correlations inherent in the low-fidelity and high-fidelity data, although, there, the focus is on accurately fitting the high-fidelity data.

Multitask learning is useful when all the training data are available; but when the data for different tasks become available over time, transfer learning is used instead. Transfer learning adapts a model learned from a library of initial training tasks to a situation in which new tasks become available. The key component of transfer learning is to identify the task similarity and the knowledge to be transferred. For instance, multiple tasks can share the same initial parameters, but once more data become available for a new task, parameter weights can be optimized according to a loss function accounting for the errors in the new task[230].

Outlook

We have discussed how AI and ML methods have transformed the computational and experimental landscape within materials research. However, for the sustenance and growth of such ML-powered materials intelligence ecosystems, several challenges should be overcome. ML methods rely on a constant flux of high-fidelity data generated in a consistent and systematic manner. However, in the absence of standard protocols for systematic and sustainable materials (meta) data capture, curation and organization, most materials databases exist in a stand-alone fashion, with the community unable to collectively exploit materials information from different channels. On the side of ML model development and testing, benchmark datasets are necessary for consistent testing of new algorithms, in keeping with common practices in computer and statistical sciences. These benchmark datasets will be a good addition to the existing culture of openly sharing ML codes and data. Understanding a ML model domain of applicability, its interpretability and its use for outlier detection still remain major challenges, which are bound to intensify as complex NN-based approaches become more popular.

Nevertheless, it is fair to claim that the notion of AI is reverberating within materials research. Early successes in building predictive models are paving the way to building experimental and computational autonomy and guided high-throughput workflows. Integration of various parts of the burgeoning materials intelligence ecosystems may lead to materials-savvy digital assistants that may intelligently and autonomously interact with both theoretical and experimental materials researchers; this human–machine partnership can lead to dramatic efficiencies, accelerated discoveries and increased productivity. From a human resources perspective, adequate training of the current and next generation of materials scientists on AI and ML methods is needed to ensure the effective and appropriate utilization of these tools.

Published online: 09 November 2020

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
2. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
3. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: Recent applications and prospects. *NPJ Comput. Mater.* **3**, 54 (2017).
4. Schmidt, J., Marques, M. R., Botti, S. & Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *NPJ Comput. Mater.* **5**, 83 (2019).
5. Coley, C. W. et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, eaax1566 (2019).
6. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **361**, 360–365 (2018).
7. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
8. The Minerals Metals & Materials Society (TMS). *Building a Materials Data Infrastructure: Opening New Pathways to Discovery and Innovation in Science and Engineering* (TMS, 2017).
9. Fisher, R. A. *The Design of Experiments* 9th edn (Macmillan, 1971).
10. Cao, B. et al. How to optimize materials and devices via design of experiments and machine learning: Demonstration using organic photovoltaics. *ACS Nano* **12**, 7434–7444 (2018).
11. Morris, M. D. & Mitchell, T. J. Exploratory designs for computational experiments. *J. Stat. Plan. Inference* **43**, 381–402 (1995).
12. Qian, P. Z. Sliced Latin hypercube designs. *J. Am. Stat. Assoc.* **107**, 393–399 (2012).
13. Joseph, V. R., Gul, E. & Ba, S. Designing computer experiments with multiple types of factors: The MaxPro approach. *J. Qual. Technol.* **52**, 343–354 (2019).
14. Zhang, Y., Yoon, H. S., Koh, C. S. & Xie, D. in *2007 International Conference on Electrical Machines and Systems (ICEMS)* 1414–1418 (IEEE, 2007).
15. Joseph, V. R. Space-filling designs for computer experiments: A review. *Qual. Eng.* **28**, 28–35 (2016).
16. Castillo, A. R. & Kalidindi, S. R. A Bayesian framework for the estimation of the single crystal elastic parameters from spherical indentation stress-strain measurements. *Front. Mater.* **6**, 136 (2019).
17. Castillo, A. R. & Kalidindi, S. R. Bayesian estimation of single ply anisotropic elastic constants from spherical indentations on multi-laminate polymer-matrix fiber-reinforced composite samples. *Meccanica* https://doi.org/10.1007/s11012-020-01154-w (2020).
18. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning* Vol. 2 (MIT Press, 2006).
19. Forrester, A. I. J., Sóbester, A. & Keane, A. J. *Engineering Design via Surrogate Modelling: A Practical Guide* (Wiley, 2008).
20. Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & De Freitas, N. Taking the human out of the loop: A review of Bayesian optimization. *Proc. IEEE* **104**, 148–175 (2015).
21. Kushner, H. J. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *J. Basic Eng.* **86**, 97–106 (1964).
22. Russo, D. J. et al. A tutorial on Thompson sampling. *Found. Trends Mach. Learn.* **11**, 1–96 (2018).
23. Xue, D. et al. Accelerated search for materials with targeted properties by adaptive design. *Nat. Commun.* **7**, 11241 (2016).
24. Kim, C., Chandrasekaran, A., Jha, A. & Ramprasad, R. Active-learning and materials design: The example of high glass transition temperature polymers. *MRS Commun.* **9**, 860–866 (2019).
25. Yuan, R. et al. Accelerated discovery of large electrostrains in $BaTiO_3$-based piezoelectrics using active learning. *Adv. Mater.* **30**, 1702884 (2018).
26. Wen, C. et al. Machine learning assisted design of high entropy alloys with desired property. *Acta Mater.* **170**, 109–117 (2019).
27. Xue, D. et al. Accelerated search for $BaTiO_3$-based piezoelectrics with vertical morphotropic phase boundary using Bayesian learning. *Proc. Natl Acad. Sci. USA* **113**, 13301–13306 (2016).
28. Lookman, T., Balachandran, P. V., Xue, D., Hogden, J. & Theiler, J. Statistical inference and adaptive design for materials discovery. *Curr. Opin. Solid State Mater. Sci.* **21**, 121–128 (2017).
29. Lookman, T., Balachandran, P. V., Xue, D. & Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *NPJ Comput. Mater.* **5**, 21 (2019).
30. Rohr, B. et al. Benchmarking the acceleration of materials discovery by sequential learning. *Chem. Sci.* **11**, 2696–2706 (2020).
31. Swain, M. C. & Cole, J. M. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.* **56**, 1894–1904 (2016).
32. Pennington, J., Socher, R. & Manning, C. D. in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 1532–1543 (Association for Computational Linguistics, 2014).
33. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. Preprint at *arXiv* https://arxiv.org/abs/1301.3781 (2013).
34. Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
35. Court, C. J. & Cole, J. M. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Sci. Data* **5**, 180111 (2018).
36. Jensen, Z. et al. A machine learning approach to zeolite synthesis enabled by automatic literature data extraction. *ACS Cent. Sci.* **5**, 892–899 (2019).
37. Kim, E. et al. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **29**, 9436–9444 (2017).
38. Kim, E. et al. Inorganic materials synthesis planning with literature-trained neural networks. *J. Chem. Inf. Model.* **60**, 1194–1201 (2020).
39. He, T. et al. Similarity of precursors in solid-state synthesis as text-mined from scientific literature. *Chem. Mater.* **32**, 7861–7873 (2020).
40. Writer, B. *Lithium-Ion Batteries. A Machine-Generated Summary of Current Research* (Springer, 2019).
41. Wu, P., Carberry, S., Elzer, S. & Chester, D. in *International Conference on Theory and Application of Diagrams* 220–234 (Springer, 2010).
42. Savva, M. et al. in *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* 393–402 (ACM, 2011).
43. Ray Choudhury, S. & Giles, C. L. in *Proceedings of the 24th International Conference on World Wide Web* 667–672 (ACM, 2015).
44. Siegel, N., Horvitz, Z., Levin, R., Divvala, S. & Farhadi, A. in *European Conference on Computer Vision* 664–680 (Springer, 2016).
45. Seo, M., Hajishirzi, H., Farhadi, A., Etzioni, O. & Malcolm, C. in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* 1466–1476 (Association for Computational Linguistics, 2015).

46. Sachan, M., Dubey, K. & Xing, E. in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* 773–784 (Association for Computational Linguistics, 2017).
47. Sachan, M. et al. Discourse in multimedia: A case study in extracting geometry knowledge from textbooks. *Comput. Linguist.* **45**, 627–665 (2019).
48. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
49. Abadi, M. et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. Preprint at *arXiv* https://arxiv.org/abs/1603.04467 (2015).
50. Mueller, T., Kusne, A. G. & Ramprasad, R. Machine learning in materials science: Recent progress and emerging applications. *Rev. Comput. Chem.* **29**, 186–273 (2016).
51. Schleder, G. R., Padilha, A. C., Acosta, C. M., Costa, M. & Fazzio, A. From DFT to machine learning: Recent approaches to materials science–a review. *J. Phys. Mater.* **2**, 032001 (2019).
52. Mannodi-Kanakkithodi, A. et al. Scoping the polymer genome: A roadmap for rational polymer dielectrics design and beyond. *Mater. Today* **21**, 785–796 (2018).
53. Huang, A., Huo, Y., Yang, J. & Li, G. Computational simulation and prediction on electrical conductivity of oxide-based melts by big data mining. *Materials* **12**, 1059 (2019).
54. Kim, C., Pilania, G. & Ramprasad, R. Machine learning assisted predictions of intrinsic dielectric breakdown strength of ABX_3 perovskites. *J. Phys. Chem. C* **120**, 14575–14580 (2016).
55. Kim, C., Pilania, G. & Ramprasad, R. From organized high-throughput data to phenomenological theory using machine learning: The example of dielectric breakdown. *Chem. Mater.* **28**, 1304–1311 (2016).
56. Santos, I., Nieves, J., Penya, Y. K. & Bringas, P. G. in *2009 ICCAS-SICE* 4536–4541 (IEEE, 2009).
57. Yaseen, Z. M. et al. Predicting compressive strength of lightweight foamed concrete using extreme learning machine model. *Adv. Eng. Softw.* **115**, 112–125 (2018).
58. De Jong, M. et al. A statistical learning framework for materials science: Application to elastic moduli of *k*-nary inorganic polycrystalline compounds. *Sci. Rep.* **6**, 34256 (2016).
59. Hamdia, K. M., Lahmer, T., Nguyen-Thoi, T. & Rabczuk, T. Predicting the fracture toughness of PNCs: A stochastic approach based on ANN and ANFIS. *Comput. Mater. Sci.* **102**, 304–313 (2015).
60. Kauwe, S. K., Graser, J., Vazquez, A. & Sparks, T. D. Machine learning prediction of heat capacity for solid inorganics. *Integrat. Mater. Manuf. Innov.* **7**, 43–51 (2018).
61. Legrain, F., Carrete, J., van Roekeghem, A., Curtarolo, S. & Mingo, N. How chemical composition alone can predict vibrational free energies and entropies of solids. *Chem. Mater.* **29**, 6220–6227 (2017).
62. Chen, L., Tran, H., Batra, R., Kim, C. & Ramprasad, R. Machine learning models for the lattice thermal conductivity prediction of inorganic materials. *Comput. Mater. Sci.* **170**, 109155 (2019).
63. Stanev, V. et al. Machine learning modeling of superconducting critical temperature. *NPJ Comput. Mater.* **4**, 29 (2018).
64. Balachandran, P. V., Kowalski, B., Sehirlioglu, A. & Lookman, T. Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning. *Nat. Commun.* **9**, 1668 (2018).
65. Zhang, Y. & Kim, E.-A. Quantum loop topography for machine learning. *Phys. Rev. Lett.* **118**, 216401 (2017).
66. Gaultois, M. W. et al. Perspective: Web-based machine learning models for real-time screening of thermoelectric materials properties. *APL Mater.* **4**, 053213 (2016).
67. Sendek, A. D. et al. Machine learning-assisted discovery of solid Li-ion conducting materials. *Chem. Mater.* **31**, 342–352 (2018).
68. Mansouri Tehrani, A. et al. Machine learning directed search for ultraincompressible, superhard materials. *J. Am. Chem. Soc.* **140**, 9844–9853 (2018).
69. Wu, Y.-J., Sasaki, M., Goto, M., Fang, L. & Xu, Y. Electrically conductive thermally insulating Bi–Si nanocomposites by interface design for thermal management. *ACS Appl. Nano Mater.* **1**, 3355–3363 (2018).
70. Ren, F. et al. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Sci. Adv.* **4**, eaaq1566 (2018).
71. Kim, C., Chandrasekaran, A., Huan, T. D., Das, D. & Ramprasad, R. Polymer genome: A data-powered polymer informatics platform for property predictions. *J. Phys. Chem. C* **122**, 17575–17585 (2018).
72. Kim, C., Batra, R., Chen, L., Tran, H. & Ramprasad, R. Polymer design using genetic algorithm and machine learning. *Comput. Mat. Sci.* **186**, 110067 (2020).
73. Yoshida, M. et al. Using evolutionary algorithms and machine learning to explore sequence space for the discovery of antimicrobial peptides. *Chem* **4**, 533–543 (2018).
74. Meredig, B. et al. Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. *Mol. Syst. Des. Eng.* **3**, 819–825 (2018).
75. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **7**, 20 (2015).
76. Venkatram, S. et al. Predicting crystallization tendency of polymers using multi-fidelity information fusion and machine learning. *J. Phys. Chem. B* **124**, 6046–6054 (2020).
77. Pilania, G., Gubernatis, J. E. & Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.* **129**, 156–163 (2017).
78. Zaspel, P., Huang, B., Harbrecht, H. & von Lilienfeld, O. A. Boosting quantum machine learning models with a multilevel combination technique: Pople diagrams revisited. *J. Chem. Theory Comput.* **15**, 1546–1559 (2018).
79. Patra, A. et al. A multi-fidelity information-fusion approach to machine learn and predict polymer bandgap. *Comput. Mater. Sci.* **172**, 109286 (2020).
80. Batra, R., Pilania, G., Uberuaga, B. P. & Ramprasad, R. Multifidelity information fusion with machine learning: A case study of dopant formation energies in hafnia. *ACS Appl. Mater. Interfaces* **11**, 24906–24918 (2019).
81. Kukreja, S. L., Löfberg, J. & Brenner, M. J. A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification. *IFAC Proc. Vol.* **39**, 814–819 (2006).
82. Ghiringhelli, L. M. et al. Learning physical descriptors for materials science by compressed sensing. *New J. Phys.* **19**, 023017 (2017).
83. Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M. & Ghiringhelli, L. M. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys. Rev. Mater.* **2**, 083802 (2018).
84. Bartel, C. J. et al. New tolerance factor to predict the stability of perovskite oxides and halides. *Sci. Adv.* **5**, eaav0693 (2019).
85. Goldschmidt, V. M. Die gesetze der krystallochemie. *Naturwissenschaften* **14**, 477–485 (1926).
86. Bartel, C. J. et al. Physical descriptor for the Gibbs energy of inorganic crystalline solids and temperature-dependent materials chemistry. *Nat. Commun.* **9**, 4168 (2018).
87. Andersen, M., Levchenko, S. V., Scheffler, M. & Reuter, K. Beyond scaling relations for the description of catalytic materials. *ACS Catal.* **9**, 2752–2759 (2019).
88. Sun, S., Ouyang, R., Zhang, B. & Zhang, T.-Y. Data-driven discovery of formulas by symbolic regression. *MRS Bull.* **44**, 559–564 (2019).
89. Wang, Y., Wagner, N. & Rondinelli, J. M. Symbolic regression in materials science. *MRS Commun.* **9**, 793–805 (2019).
90. Hernandez, A., Balasubramanian, A., Yuan, F., Mason, S. A. & Mueller, T. Fast, accurate, and transferable many-body interatomic potentials by symbolic regression. *NPJ Comput. Mater.* **5**, 112 (2019).
91. Sastry, K., Johnson, D. D., Goldberg, D. E. & Bellon, P. Genetic programming for multitimescale modeling. *Phys. Rev. B* **72**, 085438 (2005).
92. Gandomi, A. H., Sajedi, S., Kiani, B. & Huang, Q. Genetic programming for experimental big data mining: A case study on concrete creep formulation. *Autom. Constr.* **70**, 89–97 (2016).
93. Batra, R. & Sankaranarayanan, S. Machine learning for multi-fidelity scale bridging and dynamical simulations of materials. *J. Phys. Mater.* **3**, 031002 (2020).
94. Jackson, N. E., Webb, M. A. & de Pablo, J. J. Recent advances in machine learning towards multiscale soft materials design. *Curr. Opin. Chem. Eng.* **23**, 106–114 (2019).
95. Ye, W., Chen, C., Wang, Z., Chu, I.-H. & Ong, S. P. Deep neural networks for accurate predictions of crystal stability. *Nat. Commun.* **9**, 3800 (2018).
96. Jha, D. et al. ElemNet: Deep learning the chemistry of materials from only elemental composition. *Sci. Rep.* **8**, 17593 (2018).
97. DeCost, B. L., Lei, B., Francis, T. & Holm, E. A. High throughput quantitative metallography for complex microstructures using deep learning: A case study in ultrahigh carbon steel. *Microsc. Microanal.* **25**, 21–29 (2019).
98. Nash, W., Drummond, T. & Birbilis, N. A review of deep learning in the study of materials degradation. *NPJ Mater. Degrad.* **2**, 37 (2018).
99. Cecen, A., Dai, H., Yabansu, Y. C., Kalidindi, S. R. & Song, L. Material structure-property linkages using three-dimensional convolutional neural networks. *Acta Mater.* **146**, 76–84 (2018).
100. Sanyal, S. et al. MT-CGCNN: Integrating crystal graph convolutional neural network with multitask learning for material property prediction. Preprint at https://arxiv.org/abs/1811.05660 (2018).
101. Agrawal, A. & Choudhary, A. Deep materials informatics: Applications of deep learning in materials science. *MRS Commun.* **9**, 779–792 (2019).
102. Zheng, X., Zheng, P. & Zhang, R.-Z. Machine learning material properties from the periodic table using convolutional neural networks. *Chem. Sci.* **9**, 8426–8432 (2018).
103. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. Schnet–A deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).
104. Dai, H., Li, C., Coley, C., Dai, B. & Song, L. in *Advances in Neural Information Processing Systems 32* (eds Wallach, H. et al.) 8870–8880 (Curran Associates, 2019).
105. Coley, C. W. et al. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **10**, 370–377 (2019).
106. You, J., Liu, B., Ying, Z., Pande, V. & Leskovec, J. in *Advances in Neural Information Processing Systems 31* (eds Bengio, S. et al) 6410–6421 (Curran Associates, 2018).
107. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
108. Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **31**, 3564–3572 (2019).
109. Aykol, M. et al. Network analysis of synthesizable materials discovery. *Nat. Commun.* **10**, 2018 (2019).
110. Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: Moving beyond fingerprints. *J. Comput. Mol. Des.* **30**, 595–608 (2016).
111. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. Preprint at *arXiv* https://arxiv.org/abs/1312.6114 (2014).
112. Goodfellow, I. et al. in *Advances in Neural Information Processing Systems 27* (eds Ghahramani, Z. et al.) 2672–2680 (Curran Associates, 2014).
113. Li, W., Jacobs, R. & Morgan, D. Predicting the thermodynamic stability of perovskite oxides using machine learning models. *Comput. Mater. Sci.* **150**, 454–463 (2018).
114. Ziletti, A., Kumar, D., Scheffler, M. & Ghiringhelli, L. M. Insightful classification of crystal structures using deep learning. *Nat. Commun.* **9**, 2775 (2018).
115. Lennard-Jones, J. E. On the determination of molecular fields. II. From the equation of state of gas. *Proc. R. Soc. Lond. A* **106**, 463–477 (1924).
116. Chenoweth, K., Van Duin, A. C. & Goddard, W. A. ReaxFF reactive force field for molecular dynamics simulations of hydrocarbon oxidation. *J. Phys. Chem. A* **112**, 1040–1053 (2008).
117. Liu, H., Fu, Z., Li, Y., Sabri, N. F. A. & Bauchy, M. Parameterization of empirical forcefields for glassy silica using machine learning. *MRS Commun.* **9**, 593–599 (2019).
118. Chan, H. et al. Machine learning coarse grained models for water. *Nat. Commun.* **10**, 379 (2019).
119. Chan, H. et al. Machine learning a bond order potential model to study thermal transport in WSe_2 nanostructures. *Nanoscale* **11**, 10381–10392 (2019).
120. Chan, H. et al. Machine learning classical interatomic potentials for molecular dynamics from first-principles training data. *J. Phys. Chem. C* **123**, 6941–6957 (2019).

121. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
122. Deringer, V. L., Caro, M. A. & Csányi, G. Machine learning interatomic potentials as emerging tools for materials science. *Adv. Mater.* **31**, 1902765 (2019).
123. Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **145**, 170901 (2016).
124. Handley, C. M. & Popelier, P. L. Potential energy surfaces fitted by artificial neural networks. *J. Phys. Chem. A* **114**, 3371–3383 (2010).
125. Botu, V., Batra, R., Chapman, J. & Ramprasad, R. Machine learning force fields: Construction, validation, and outlook. *J. Phys. Chem. C* **121**, 511–522 (2017).
126. Huan, T. D. et al. A universal strategy for the creation of machine learning-based atomistic force fields. *NPJ Comput. Mater.* **3**, 37 (2017).
127. Rowe, P., Csányi, G., Alfè, D. & Michaelides, A. Development of a machine learning potential for graphene. *Phys. Rev. B* **97**, 054303 (2018).
128. Podryabinkin, E. V., Tikhonov, E. V., Shapeev, A. V. & Oganov, A. R. Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning. *Phys. Rev. B* **99**, 064114 (2019).
129. Podryabinkin, E. V. & Shapeev, A. V. Active learning of linearly parametrized interatomic potentials. *Comput. Mater. Sci.* **140**, 171–180 (2017).
130. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
131. Deringer, V. L. et al. Realistic atomistic structure of amorphous silicon from machine-learning-driven molecular dynamics. *J. Phys. Chem. Lett.* **9**, 2879–2885 (2018).
132. Dragoni, D., Daff, T. D., Csányi, G. & Marzari, N. Achieving DFT accuracy with a machine-learning interatomic potential: Thermomechanics and defects in bcc ferromagnetic iron. *Phys. Rev. Mater.* **2**, 013808 (2018).
133. Zong, H., Pilania, G., Ding, X., Ackland, G. J. & Lookman, T. Developing an interatomic potential for martensitic phase transformations in zirconium by machine learning. *NPJ Comput. Mater.* **4**, 48 (2018).
134. Artrith, N. & Kolpak, A. M. Grand canonical molecular dynamics simulations of Cu–Au nanoalloys in thermal equilibrium using reactive ANN potentials. *Comput. Mater. Sci.* **110**, 20–28 (2015).
135. Chiriki, S. & Bulusu, S. S. Modeling of DFT quality neural network potential for sodium clusters: Application to melting of sodium clusters (Na20 to Na40). *Chem. Phys. Lett.* **652**, 130–135 (2016).
136. Chiriki, S., Jindal, S. & Bulusu, S. S. Neural network potentials for dynamics and thermodynamics of gold nanoparticles. *J. Chem. Phys.* **146**, 084314 (2017).
137. Sosso, G. C., Miceli, G., Caravati, S., Behler, J. & Bernasconi, M. Neural network interatomic potential for the phase change material GeTe. *Phys. Rev. B* **85**, 174103 (2012).
138. Artrith, N., Morawietz, T. & Behler, J. High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide. *Phys. Rev. B* **83**, 153101 (2011).
139. Artrith, N. & Urban, A. An implementation of artificial neural-network potentials for atomistic materials simulations: Performance for TiO$_2$. *Comput. Mater. Sci.* **114**, 135–150 (2016).
140. Morawietz, T., Singraber, A., Dellago, C. & Behler, J. How van der Waals interactions determine the unique properties of water. *Proc. Natl Acad. Sci. USA* **113**, 8368–8373 (2016).
141. Cheng, B., Behler, J. & Ceriotti, M. Nuclear quantum effects in water at the triple point: Using theory as a link between experiments. *J. Phys. Chem. Lett.* **7**, 2210–2215 (2016).
142. Jose, K. J., Artrith, N. & Behler, J. Construction of high-dimensional neural network potentials using environment-dependent atom pairs. *J. Chem. Phys.* **136**, 194111 (2012).
143. Gastegger, M., Kauffmann, C., Behler, J. & Marquetand, P. Comparing the accuracy of high-dimensional neural network potentials and the systematic molecular fragmentation method: A benchmark study for all-trans alkanes. *J. Chem. Phys.* **144**, 194110 (2016).
144. Boes, J. R. & Kitchin, J. R. Neural network predictions of oxygen interactions on a dynamic Pd surface. *Mol. Simul.* **43**, 346–354 (2017).
145. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
146. Khorshidi, A. & Peterson, A. A. Amp: A modular approach to machine learning in atomistic simulations. *Comput. Phys. Commun.* **207**, 310–324 (2016).
147. Wang, H., Zhang, L., Han, J. & Weinan, E. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Comput. Phys. Commun.* **228**, 178–184 (2018).
148. Shapeev, A. V. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Model. Simul.* **14**, 1153–1173 (2016).
149. Schutt, K. et al. SchNetPack: A deep learning toolbox for atomistic systems. *J. Chem. Theory Comput.* **15**, 448–455 (2018).
150. Desai, S., Reeve, S. T. & Belak, J. F. Implementing a neural network interatomic model with performance portability for emerging exascale architectures. Preprint at *arXiv* https://arxiv.org/abs/2002.00054 (2020).
151. Thompson, A. P., Swiler, L. P., Trott, C. R., Foiles, S. M. & Tucker, G. J. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* **285**, 316–330 (2015).
152. Zuo, Y. et al. Performance and cost assessment of machine learning interatomic potentials. *J. Phys. Chem. A* **124**, 731–745 (2020).
153. Artrith, N., Urban, A. & Ceder, G. Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species. *Phys. Rev. B* **96**, 014112 (2017).
154. Musil, F., Willatt, M. J., Langovoy, M. A. & Ceriotti, M. Fast and accurate uncertainty estimation in chemical machine learning. *J. Chem. Theory Comput.* **15**, 906–915 (2019).
155. Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O. & Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **148**, 241733 (2018).
156. Huan, T. D. et al. Iterative-learning strategy for the development of application-specific atomistic force fields. *J. Phys. Chem. C* **123**, 20715–20722 (2019).
157. Smith, J. S. et al. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **10**, 2903 (2019).
158. Brockherde, F. et al. Bypassing the Kohn-Sham equations with machine learning. *Nat. Commun.* **8**, 872 (2017).
159. Grisafi, A. et al. Transferable machine-learning model of the electron density. *ACS Cent. Sci.* **5**, 57–64 (2018).
160. Chandrasekaran, A. et al. Solving the electronic structure problem with machine learning. *NPJ Comput. Mater.* **5**, 22 (2019).
161. Kamal, D., Chandrasekaran, A., Batra, R. & Ramprasad, R. A charge density prediction model for hydrocarbons using deep neural networks. *Mach. Learn.* **1**, 025003 (2020).
162. Schmidt, J., Benavides-Riveros, C. L. & Marques, M. A. Machine learning the physical nonlocal exchange–correlation functional of density-functional theory. *J. Phys. Chem. Lett.* **10**, 6425–6431 (2019).
163. Lei, X. & Medford, A. J. Design and analysis of machine learning exchange-correlation functionals via rotationally invariant convolutional descriptors. *Phys. Rev. Mater.* **3**, 063801 (2019).
164. Snyder, J. C., Rupp, M., Hansen, K., Müller, K.-R. & Burke, K. Finding density functionals with machine learning. *Phys. Rev. Lett.* **108**, 253002 (2012).
165. Comer, M., Bouman, C. A., De Graef, M. & Simmons, J. P. Bayesian methods for image segmentation. *JOM* **63**, 55–57 (2011).
166. Simmons, J. et al. Application and further development of advanced image processing algorithms for automated analysis of serial section image data. *Model. Simul. Mater. Sci. Eng.* **17**, 025002 (2008).
167. Gibert, X., Patel, V. M. & Chellappa, R. Deep multitask learning for railway track inspection. *IEEE Trans. Intell. Transp Syst.* **18**, 153–164 (2016).
168. Niezgoda, S. R., Yabansu, Y. C. & Kalidindi, S. R. Understanding and visualizing microstructure and microstructure variance as a stochastic process. *Acta Mater.* **59**, 6387–6400 (2011).
169. Steinmetz, P. et al. Analytics for microstructure datasets produced by phase-field simulations. *Acta Mater.* **103**, 192–203 (2016).
170. Yang, Z. et al. Establishing structure-property localization linkages for elastic deformation of three-dimensional high contrast composites using deep learning approaches. *Acta Mater.* **166**, 335–345 (2019).
171. Yang, Z. et al. Deep learning approaches for mining structure-property linkages in high contrast composites from simulation datasets. *Comput. Mater. Sci.* **151**, 278–287 (2018).
172. Yang, Z. et al. Microstructural materials design via deep adversarial learning methodology. *J. Mech. Des.* **140**, 111416 (2018).
173. Jha, D. et al. Extracting grain orientations from EBSD patterns of polycrystalline materials using convolutional neural networks. *Microsc. Microanal.* **24**, 497–502 (2018).
174. Ziatdinov, M. et al. Imaging mechanism for hyperspectral scanning probe microscopy via Gaussian process modelling. *NPJ Comput. Mater.* **6**, 21 (2020).
175. Maksov, A. et al. Deep learning analysis of defect and phase evolution during electron beam-induced transformations in WS$_2$. *NPJ Comput. Mater.* **5**, 12 (2019).
176. Vasudevan, R. K. et al. Mapping mesoscopic phase evolution during E-beam induced transformations via deep learning of atomically resolved images. *NPJ Comput. Mater.* **4**, 30 (2018).
177. Laanait, N., He, Q. & Borisevich, A. Y. Reconstruction of 3-D atomic distortions from electron microscopy with deep learning. Preprint at *arXiv* https://arxiv.org/abs/1902.06876 (2019).
178. Laanait, N., Yin, J. & Borisevich, A. in *Conference on Neural Information Processing Systems (NeurIPS) 2019 Workshop Deep Inverse* (OpenReview, 2019).
179. Cherukara, M. J., Nashed, Y. S. & Harder, R. J. Real-time coherent diffraction inversion using deep generative networks. *Sci. Rep.* **8**, 16520 (2018).
180. Godaliyadda, G. D. et al. A supervised learning approach for dynamic sampling. *Electron. Imaging* **2016**, 1–8 (2016).
181. Attia, P. M. et al. Closed-loop optimization of fast-charging protocols for batteries with machine learning. *Nature* **578**, 397–402 (2020).
182. Severson, K. A. et al. Data-driven prediction of battery cycle life before capacity degradation. *Nat. Energy* **4**, 383–391 (2019).
183. Kusne, A. G. et al. On-the-fly machine-learning for high-throughput experiments: Search for rare-earth-free permanent magnets. *Sci. Rep.* **4**, 6367 (2014).
184. Long, C. et al. Rapid structural mapping of ternary metallic alloy systems using the combinatorial approach and cluster analysis. *Rev. Sci. Instrum.* **78**, 072217 (2007).
185. Long, C., Bunker, D., Li, X., Karen, V. & Takeuchi, I. Rapid identification of structural phases in combinatorial thin-film libraries using x-ray diffraction and non-negative matrix factorization. *Rev. Sci. Instrum.* **80**, 103902 (2009).
186. Suram, S. K. et al. Automated phase mapping with AgileFD and its application to light absorber discovery in the V–Mn–Nb oxide system. *ACS Comb. Sci.* **19**, 37–46 (2017).
187. Gomes, C. P. et al. CRYSTAL: a multi-agent AI system for automated mapping of materials' crystal structures. *MRS Commun.* **9**, 600–608 (2019).
188. Bai, J. et al. Phase mapper: Accelerating materials discovery with AI. *AI Mag.* **39**, 15–26 (2018).
189. Tabor, D. P. et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* **3**, 5–20 (2018).
190. King, R. D. et al. The automation of science. *Science* **324**, 85–89 (2009).
191. Williams, K. et al. Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. *J. R. Soc. Interface* **12**, 20141289 (2015).
192. Nikolaev, P. et al. Autonomy in materials research: A case study in carbon nanotube growth. *NPJ Comput. Mater.* **2**, 16031 (2016).
193. Wigley, P. B. et al. Fast machine-learning online optimization of ultra-cold-atom experiments. *Sci. Rep.* **6**, 25890 (2016).
194. Duros, V. et al. Human versus robots in the discovery and crystallization of gigantic polyoxometalates. *Angew. Chem.* **129**, 10955–10960 (2017).
195. Noack, M. M. et al. A kriging-based approach to autonomous experimentation with applications to x-ray scattering. *Sci. Rep.* **9**, 11809 (2019).
196. Masubuchi, S. et al. Autonomous robotic searching and assembly of two-dimensional crystals to build van der Waals superlattices. *Nat. Commun.* **9**, 1413 (2018).

197. Chen, S. et al. Exploring the stability of novel wide bandgap perovskites by a robot based high throughput approach. *Adv. Energy Mater.* **8**, 1701543 (2018).
198. Jensen, K. F. Automated synthesis on a hub-and-spoke system. *Nature* **579**, 346–348 (2020).
199. Roch, L. M. et al. Chemos: An orchestration software to democratize autonomous discovery. *PLoS ONE* **15**, e0229862 (2020).
200. Montoya, J. H. et al. Autonomous intelligent agents for accelerated materials discovery. *Chem. Sci.* **11**, 8517–8532 (2020).
201. Mannodi-Kanakkithodi, A. et al. Rational co-design of polymer dielectrics for energy storage. *Adv. Mater.* **28**, 6277–6291 (2016).
202. Mannodi-Kanakkithodi, A., Pilania, G., Huan, T. D., Lookman, T. & Ramprasad, R. Machine learning strategy for accelerated design of polymer dielectrics. *Sci. Rep.* **6**, 20952 (2016).
203. Pilania, G., Iverson, C. N., Lookman, T. & Marrone, B. L. Machine-learning-based predictive modeling of glass transition temperatures: A case of polyhydroxyalkanoate homopolymers and copolymers. *J. Chem. Inf. Model.* **59**, 5013–5025 (2019).
204. Batra, R. et al. Polymers for extreme conditions designed using syntax-directed variational autoencoders. Preprint at http://arxiv.org/abs/2011.02551v1 (2020).
205. Dai, H., Tian, Y., Dai, B., Skiena, S. & Song, L. Syntax-directed variational autoencoder for structured data. Preprint at *arXiv* https://arxiv.org/abs/1802.08786 (2018).
206. Kim, B., Lee, S. & Kim, J. Inverse design of porous materials using artificial neural networks. *Sci. Adv.* **6**, eaax9324 (2020).
207. Corey, E. J. *The Logic of Chemical Synthesis* (Wiley, 1991).
208. Coley, C. W., Green, W. H. & Jensen, K. F. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* **51**, 1281–1289 (2018).
209. Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H. & Jensen, K. F. Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci.* **3**, 434–443 (2017).
210. Segler, M. H., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–601 (2018).
211. Jin, W., Coley, C., Barzilay, R. & Jaakkola, T. Predicting organic reaction outcomes with Weisfeiler-Lehman network. in *Advances in Neural Information Processing Systems* 2607–2616 (Cornell University, 2017).
212. Bradshaw, J., Kusner, M. J., Paige, B., Segler, M. H. & Hernández-Lobato, J. M. A generative model for electron paths. Preprint at *arXiv* https://arxiv.org/abs/1805.10970 (2018).
213. Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. Computer-assisted retrosynthesis based on molecular similarity. *ACS Cent. Sci.* **3**, 1237–1245 (2017).
214. Segler, M. H. & Waller, M. P. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry* **23**, 5966–5971 (2017).
215. Liu, B. et al. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.* **3**, 1103–1113 (2017).
216. Karpov, P., Godin, G. & Tetko, I. V. in *International Conference on Artificial Neural Networks* 817–830 (Springer, 2019).
217. Schwaller, P., Gaudin, T., Lanyi, D., Bekas, C. & Laino, T. "Found in Translation": Predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **9**, 6091–6098 (2018).
218. Schwaller, P. et al. Molecular transformer for chemical reaction prediction and uncertainty estimation. *ACS Cent. Sci.* **5**, 1572–1583 (2018).
219. Kim, E., Huang, K., Jegelka, S. & Olivetti, E. Virtual screening of inorganic materials synthesis parameters with deep learning. *NPJ Comput. Mater.* **3**, 53 (2017).
220. Rossi, F., Van Beek, P. & Walsh, T. *Handbook of Constraint Programming* (Elsevier, 2006).
221. Pun, G. P., Batra, R., Ramprasad, R. & Mishin, Y. Physically informed artificial neural networks for atomistic modeling of materials. *Nat. Commun.* **10**, 2339 (2019).
222. Zaheer, M. et al. in *Advances in Neural Information Processing Systems* 3391–3401 (Cornell University, 2017).
223. Schütt, K. et al. in *Advances in Neural Information Processing Systems* 991–1001 (Cornell University, 2017).
224. Zhang, L. et al. in *Advances in Neural Information Processing Systems* 4436–4446 (Association for Computing Machinery, 2018).
225. Noé, F., Tkatchenko, A., Müller, K.-R. & Clementi, C. Machine learning for molecular simulation. *Annu. Rev. Phys. Chem.* **71**, 361–390 (2020).
226. Chmiela, S. et al. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, e1603015 (2017).
227. Chandrasekaran, A., Kim, C., Venkatram, S. & Ramprasad, R. A deep learning solvent-selection paradigm powered by a massive solvent/nonsolvent database for polymers. *Macromolecules* **53**, 4764–4769 (2020).
228. Zhu, G. et al. Polymer genome–based prediction of gas permeabilities in polymers. *J. Polym. Eng.* **40**, 451–457 (2020).
229. Zubatyuk, R., Smith, J. S., Leszczynski, J. & Isayev, O. Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Sci. Adv.* **5**, eaav6490 (2019).
230. Finn, C., Abbeel, P. & Levine, S. in *Proceedings of the 34th International Conference on Machine Learning - Volume 70* 1126–1135 (JMLR.org, 2017).
231. Levin, I. *NIST Inorganic Crystal Structure Database (ICSD)* (National Institute of Standards and Technology, 2018).
232. Pauling File. paulingfile.com (2020).
233. Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y. & Yamazaki, M. in *2011 International Conference on Emerging Intelligent Data and Web Technologies* 22–29 (IEEE, 2011).
234. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge structural database. *Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater.* **72**, 171–179 (2016).
235. MatWeb. www.matweb.com (2020).
236. Total Materia. www.totalmateria.com (2020).
237. INTERGLAD. www.newglass.jp/interglad_n/gaiyo/info_e.html (2020).
238. Mindat. www.mindat.org (2020).
239. ASM International. www.asminternational.org (2020).
240. Downs, R. T. & Hall-Wallace, M. The *American Mineralogist* crystal structure database. *Am. Mineral* **88**, 247–250 (2003).
241. O'Mara, J., Meredig, B. & Michel, K. Materials data infrastructure: A case study of the citrination platform to examine data import, storage, and access. *JOM* **68**, 2031–2034 (2016).
242. Zagorac, D., Müller, H., Ruehl, S., Zagorac, J. & Rehme, S. Recent developments in the Inorganic Crystal Structure Database: Theoretical crystal structure data and related features. *J. Appl. Crystallogr.* **52**, 918–925 (2019).
243. Pence, H. E. & Williams, A. ChemSpider: An online chemical information resource. *J. Chem. Educ.* **87**, 1123–1124 (2010).
244. Ogata, T. & Yamazaki, M. in *Harnessing The Materials Genome: Accelerated Materials Development via Computational and Experimental Tools*, *ECI Symposium Series* (ECI Digital Archives, 2012).
245. NIST Materials Data Repository. materialsdata.nist.gov (2020).
246. Zhao, H. et al. Perspective: NanoMine: A material genome approach for polymer nanocomposites analysis and design. *APL Mater.* **4**, 053204 (2016).
247. SpringerMaterials Databases. materials.springer.com (2020).
248. Quirós, M., Gražulis, S., Girdzijauskaite˙, S., Merkys, A. & Vaitkus, A. Using SMILES strings for the description of chemical connectivity in the Crystallography Open Database. *J. Cheminform.* **10**, 23 (2018).
249. Jain, A. et al. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
250. Kirklin, S. et al. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *NPJ Comput. Mater.* **1**, 15010 (2015).
251. Calderon, C. E. et al. The AFLOW standard for high-throughput materials science calculations. *Comput. Mater. Sci.* **108**, 233–238 (2015).
252. Choudhary, K. et al. Computational screening of high-performance optoelectronic materials using OptB88vdW and TB-mBJ formalisms. *Sci. Data* **5**, 180082 (2018).
253. Hafiz, H. et al. A high-throughput data analysis and materials discovery tool for strongly correlated materials. *NPJ Comput. Mater.* **4**, 63 (2018).
254. Hummelshøj, J. S., Abild-Pedersen, F., Studt, F., Bligaard, T. & Nørskov, J. K. CatApp: A web application for surface chemistry and heterogeneous catalysis. *Angew. Chem. Int. Ed.* **51**, 272–274 (2012).
255. NOMAD Centre of Excellence. nomad-coe.eu (2020).
256. Nieves, P. et al. Database of novel magnetic materials for high-performance permanent magnet development. *Comput. Mater. Sci.* **168**, 188–202 (2019).
257. Spencer, P. A brief history of CALPHAD. *Calphad* **32**, 1–8 (2008).
258. Landis, D. D. et al. The computational materials repository. *Comput. Sci. Eng.* **14**, 51–57 (2012).
259. Ashton, M., Paul, J., Sinnott, S. B. & Hennig, R. G. Topology-scaling identification of layered solids and stable exfoliated 2D materials. *Phys. Rev. Lett.* **118**, 106101 (2017).

Acknowledgements

R.R. is grateful to the Office of Naval Research, the Toyota Research Institute, the Department of Energy and the National Science Foundation for financial support on machine-learning-related research through several grants. R.B was supported by Laboratory Directed Research and Development (LDRD) funding from Argonne National Laboratory, provided by the Director, Office of Science, of the U.S. Department of Energy under contract no. DE-AC02-06CH11357. Use of the Center for Nanoscale Materials, an Office of Science user facility, was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under contract no. DE-AC02-06CH11357. Discussions with K. Lipkowitz on various aspects of informatics and machine learning are greatly acknowledged. The authors are thankful to B. Storey for critical comments on the manuscript.

Author contributions

The authors contributed equally to all aspects of the article.

Competing interests

The authors declare no competing interests.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.